# NONPARAMETRIC ESTIMATION OF ECONOMETRIC MODELS WITH CATEGORICAL VARIABLES

A Dissertation

by

DESHENG OUYANG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2005

Major Subject: Economics

# NONPARAMETRIC ESTIMATION OF ECONOMETRIC MODELS WITH CATEGORICAL VARIABLES

A Dissertation

by

DESHENG OUYANG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

| | |
|---|---|
| Chair of Committee, | Qi Li |
| Committee Members, | Badi Baltagi |
| | Dennis Jansen |
| | Joel Zinn |
| Head of Department, | Leonardo Auernheimer |

August 2005

Major Subject: Economics

# ABSTRACT

Nonparametric Estimation of Econometric Models With Categorical Variables.

(August 2005)

Desheng Ouyang, B.S., Huazhong University of Science and Technology;

M.S., Huazhong University of Science and Technology

Chair of Advisory Committee: Dr. Qi Li

In this dissertation I investigate several topics in the field of nonparametric econometrics.

In chapter II, we consider the problem of estimating a nonparametric regression model with only categorical regressors. We investigate the theoretical properties of least squares cross-validated smoothing parameter selection, establish the rate of convergence (to zero) of the smoothing parameters for relevant regressors, and show that there is a high probability that the smoothing parameters for irrelevant regressors converge to their upper bound values thereby smoothing out the irrelevant regressors.

In chapter III, we consider the problem of estimating a joint distribution defined over a set of discrete variables. We use a smoothing kernel estimator to estimate the joint distribution, allowing for the case in which some of the discrete variables are uniformly distributed, and explicitly address the vector-valued smoothing parameter case due to its practical relevance. We show that the cross-validated smoothing parameters differ in their asymptotic behavior depending on whether a variable is uniformly distributed or not.

In chapter IV, we consider a k-n-n estimation of regression function with $k$ selected by a cross validation method. We consider both the local constant and local

linear cases. In both cases, the convergence rate of of the cross validated $k$ is established.

In chapter V, we consider nonparametric estimation of regression functions with mixed categorical and continuous data. The smoothing parameters in the model are selected by a cross-validation method. The uniform convergence rate of the kernel regression function estimator function with weakly dependent data is derived.

To My Mother

# ACKNOWLEDGMENTS

I thank Dr. Qi Li for his valuable advice and for encouraging me to finish this dissertation as well as Dr. Jeff Racine for his help. I also thank Dr. Badi Baltagi, Dr. Dennis Jansen, Dr. Joel Zinn for their helpful advice and comments.

TABLE OF CONTENTS

CHAPTER Page

# LIST OF TABLES

# CHAPTER I

## INTRODUCTION

The seminal work of Aitchison and Aitken (1976) has given rise to a rich literature on the kernel smoothing of discrete (categorical) variables. This literature was motivated mainly by the need to deal with the 'small cell problem' frequently encountered in the analysis of multivariate discrete data. A survey of this literature leads one rather quickly to the realization that concern lies almost exclusively with the estimation of the (conditional) probability distribution of the discrete variables. Much less effort has been paid to the regression framework when dealing with discrete regressors.

In chapter II, we study the theoretical properties of a data-driven least squares cross-validation method for selecting the smoothing parameters in a regression model when all regressors are discrete. We consider a general nonparametric regression model in which we allow for the possibility that some of the discrete variables have a natural ordering, for example, preferences (dislike, indifference, like), health (excellent, good, poor), and so forth. We derive the rate of convergence of the cross-validated smoothing parameters associated with the relevant regressors. We also demonstrate theoretically that, when irrelevant regressors are present, the associated smoothing parameters do not converge to zero, rather, with high probability they converge to their upper bound values thereby smoothing out irrelevant regressors. Finally, we provide two illustrative applications which clearly demonstrate that our nonparametric approach can produce superior out-of-sample predictions relative to those generated by some popular parametric estimation methods.

Recently, Li and Racine (2004), Hall, Racine and Li (2004), Hall, Li and Racine

---

The journal model is *IEEE Transactions on Automatic Control*.

(2004), and Racine and Li (2004) have considered nonparametric estimation of regression function and conditional density function with mixed discrete and continuous variables. The result of chapter II shows that regression model with only discrete variables differs significantly from the mixed discrete and continuous variables case. In the mixed variable case with at least one relevant continuous regressor, the irrelevant variables can be smoothed out with probability approaching one as sample size increases, in the discrete regressor only case, while the irrelevant variables can be smoothed out with a high probability, the probability is strictly less than one even as the sample size goes to infinity. Also, for the discrete regressor only case, the smoothing parameters associated with the relevant variables converge to 0 at the rate of $n^{-1}$ or $n^{-1/2}$, depending on whether there exists some irrelevant variables or not. This results again differs from the mixed variable case where the presence of irrelevant variables do no affect the rate of convergence of the smoothing parameters associated with the relevant variables.

It is well known that the traditional 'frequency-based' nonparametric approach may be used to consistently estimate a joint probability distribution defined over categorical variables. In chapter III, we focus our attention on an alternative nonparametric approach when confronted with discrete variables where, instead of using the conventional 'frequency' method to handle the discrete variables, we choose to *smooth* the discrete variables.

From a statistical perspective, the kernel smoothing of discrete variables may introduce some finite sample bias; however, it also reduce the variance, leading to a reduction in the mean square error of the resulting estimator relative to the frequency-based estimator. More importantly, we show that the kernel smoothing method, particularly when used in conjunction with a data-driven method of bandwidth selection such as cross-validation, has the ability to smooth over the 'uniformly distributed'

variables (a specific definition of 'uniformly distributed variables' is given in Section B of chapter III). When uniformly distributed variables are encountered, the conventional frequency method still splits the sample into many discrete cells, including those cells arising from the presence of the uniformly distributed variables. In contrast, the smoothing-cross-validation method can automatically oversmooth these variables with a high probability, thereby reducing the dimension of the nonparametric model to that associated with the non-uniformly distributed variables only. In such cases, we reduce estimation variance substantially without increasing bias, and we often obtain impressive efficiency gains.

As we will show in section C of chapter III, the asymptotic behavior of the smoothing parameters depends on whether a variable is uniformly distributed or not. We emphasize here that it is important to explicitly consider the vector-valued smoothing parameter case. If one were to use a scalar smoothing parameter (i.e.,the same value for each variable), as is often done for the convenience of the theoretical analyses (e.g., Li and Racine (2003)), then it would not be possible to benefit from the differing asymptotic behavior alluded to above. Thus, the results here extend Racine and Li (2003) to the practically relevant vector-valued smoothing parameter case, allowing for the existence of uniformly distributed variables, which results in differing asymptotic behavior of the smoothing parameters.

It is well known that nonparametric estimation techniques have the advantage of being robust to functional form specifications. Nonparametric kernel method, series method and k nearest neighbor (k-nn) method are all widely used by applied econometricians. It is also well established that the selection of smoothing parameters are of crucial importance in nonparametric estimations. In nonparametric regression model estimation based on kernel or series techniques, various data-driven methods are developed. With independent data, the least squared leave-one-out cross-

validation method based on the kernel or series estimators is the most popular choice of selecting the smoothing parameters. Intuitively, one can also apply the leave-one-out cross-validation method to selecting the smoothing parameter when using the k-nn nonparametric estimation method. However, to the best of out knowledge, the asymptotic behavior of the cross-validation selected smoothing parameter in a k-nn framework is not available in the literature. In chapter IVwe consider both a local constant and a local linear k-nn estimator and we provide asymptotic analysis for the cross-validation selected $k$ in both local constant and local linear k-nn regression function estimations.

Different nonparametric estimation methods have their own advantages in different situations. Undoubtedly k-nn method might be the preferred method in some applications. For example, when data points are unevenly distributed in its support, as is often the case for nonexperimental data in social sciences, kernel method with a fixed smoothing parameter (over the whole data range) may contain few data points in certain range of the support and thus may lead to unreliable estimation and inference results. In this case a researcher may prefer using a k-nn method which always uses (the nearest) $k$ data points in the nonparametric estimation. Although we will only consider the independent data case in this paper, the results of the paper can be extended to the weakly dependent case (say $\beta$ or $\alpha$ mixing processes) in a straightforward way.

It is well known that the traditional 'frequency-based' approach can be used to consistently estimate a nonparametric regression function with categorical variables. However, the use of this conventional frequency-based approach to handle the discrete variables is known to be unsatisfactory. Because many economic data contain a mixture of discrete and continuous variables, and when the sample size is not sufficiently large compared with the number of discrete cells, the frequency-based method cannot

lead to accurate nonparametric estimation results. In a seminal paper Aitchison and Aitken (1976) propose an novel idea of smoothing the discrete variables. Recently, Hall, Racine and Li (2004), Li and Racine (2003), and Racine and Li (2004) have extended the idea of Aitchison and Aitken (1976) to the case with mixed discrete and continuous variables. They suggest to smooth both the discrete and continuous variables, and use the data-driven cross-validation method to select the smoothing parameters. Hall, Racine and Li (2004) have shown that the cross-validation method has the amazing ability of (asymptotically) automatically removing irrelevant variables. This is important for nonparametric estimation, since the 'curse of dimensionality' is the main obstacle of nonparametric estimation method.

The aim of chapter V is to establish uniform convergence rates of nonparametric regression or density estimators in the presence of mixed discrete and continuous variables, we allow for deterministic or data-driven selected smoothing parameters. This is particularly important since data-driven method seems to be the only practical method to select the smoothing parameters in the mixed *categorical* and continuous variable case.

# CHAPTER II

# NONPARAMETRIC ESTIMATION OF REGRESSION FUNCTIONS WITH DISCRETE EXPLANATORY VARIABLES

## A. Introduction

The seminal work of Aitchison and Aitken (1976) has given rise to a rich literature on the kernel smoothing of discrete (categorical) variables. This literature was motivated mainly by the need to deal with the 'small cell problem' frequently encountered in the analysis of multivariate discrete data. Examples of this literature would include the work of Titterington (1980), Hall (1981), Wang and Ryzin (1981), Bierens (1983), Bowman, Hall and Titterington (1984), Grund (1993), Grund and Hall (1993), to mention only a few (see also the monographs by Scott (1992), Fahrmeir and Tutz (1994), Simonoff (1996)). A survey of this literature leads one rather quickly to the realization that concern lies almost exclusively with the estimation of the (conditional) probability distribution of the discrete variables. Much less effort has been paid to the regression framework when dealing with discrete regressors.

In this chapter we study the theoretical properties of a data-driven least squares cross-validation method for selecting the smoothing parameters in a regression model when all regressors are discrete. Our analysis is more complex than the probability distribution framework due to the existence of the random denominator in the nonparametric kernel estimator. We consider a general nonparametric regression model in which we allow for the possibility that some of the discrete variables have a natural ordering, for example, preferences (dislike, indifference, like), health (excellent, good, poor), and so forth. We derive the rate of convergence of the cross-validated smoothing parameters associated with the relevant regressors. We also demonstrate

theoretically that, when irrelevant regressors are present, the associated smoothing parameters do not converge to zero, rather, with high probability they converge to their upper bound values thereby smoothing out irrelevant regressors. A small scale simulation shows that our cross-validation to examine the finite sample performance Finally, we provide two illustrative applications which clearly demonstrate that our nonparametric approach can produce superior out-of-sample predictions relative to those generated by some popular parametric estimation methods which have been used to model these datasets, while the method is seen to remove a number of 'irrelevant' predictive variables for each example.

Recently, Li and Racine (2004), Hall, Racine and Li (2004), Hall, Li and Racine (2004), and Racine and Li (2004) have considered nonparametric estimation of regression function and conditional density function with mixed discrete and continuous variables. The result of this chapter shows that regression model with only discrete variables differs significantly from the mixed discrete and continuous variables case. In the mixed variable case with at least one relevant continuous regressor, the irrelevant variables can be smoothed out with probability approaching one as sample size increases, in the discrete regressor only case, while the irrelevant variables can be smoothed out with a high probability, the probability is strictly less than one even as the sample size goes to infinity. Also, for the discrete regressor only case, the smoothing parameters associated with the relevant variables converge to 0 at the rate of $n^{-1}$ or $n^{-1/2}$, depending on whether there exists some irrelevant variables or not. This results again differs from the mixed variable case where the presence of irrelevant variables do no affect the rate of convergence of the smoothing parameters associated with the relevant variables. Therefore, the discrete regressor only model needs a separate treatment since the results cannot be obtained as a special case from a mixed discrete and continuous variable model.

## B. Kernel Regression with Discrete Regressors

### 1. Regression with Irrelevant Regressors

Consider a nonparametric regression model given by

$$Y_i = g(X_i) + u_i, \tag{2.1}$$

where $g(\cdot)$ is an unknown function, $X_i$ is a $r \times 1$ vector of discrete variables, and $u_i$ is an error term satisfying $E(u_i|X_i) = 0$.

We allow for the possibility that some of the regressors are irrelevant in the sense that they are in fact independent of $Y_i$. Without loss of generality we assume that the first $r_1$ $(0 < r_1 \leq r)$ components of $X_i$ have relevant variables, while the remaining $r_2 = r - r_1$ components of $X_i$ are irrelevant ones. Let $\bar{X}_i$ denote the $r_1$-dimensional relevant components of $X_i$, and $\tilde{X}_i$ denote the $r_2$-dimensional irrelevant components. We ask that

$$(Y, \bar{X}) \text{ and } \tilde{X} \text{ are independent with each other.} \tag{2.2}$$

We use $x_t$ to denote the $t$-th component of $x$, and we assume that $x_t$ takes $c_t$ different values $(c_t \geq 2, t = 1, \ldots, k)$. For expositional simplicity we will only consider the case that $x$ are un-ordered discrete variables,[1] and postpone the treatment of ordered discrete variables case to the end of this section.

For an unordered variable, we suggest using a variation of Aitchison and Aitken's kernel function:

$$l(X_{i,t}, x_t, \lambda_t) = \begin{cases} 1, & \text{when } X_{i,t} = x_t, \\ \lambda_t, & \text{otherwise.} \end{cases} \tag{2.3}$$

Note that $\lambda_t = 0$ leads to an indicator function, and $\lambda_t = 1$ gives an uniform

[1]Examples of unordered discrete variables would include different regions, blood types, and so on.

weight function. Therefore, the range of $\lambda$ is $[0, 1]$ for all $t = 1, \ldots, r$.

Let $\mathbf{1}(A)$ denote the usual indicator function, which assumes the value 1 if $A$ holds true, and 0 otherwise. Combining (5.4) and (5.3), we obtain the product kernel function given by

$$L(X_i, x, \lambda_t) = \prod_{t=1}^{r} \lambda_t^{|X_{i,t} - x_t|}. \tag{2.4}$$

Observe that the kernel weight function we use does not add up to 1 when $\lambda_t \neq 0$, however, this does not affect the nonparametric estimator $\hat{g}(x)$ defined in equation (2.6) below as the kernel function appears in both the numerator and the denominator of Equation (2.6), thus the kernel function can be multiplied by any non-zero constant leaving the definition of $\hat{g}(x)$ intact.

We use $\mathcal{D}$ to denote the range assumed by $X_i$. For $x \in \mathcal{D}$, we estimate the probability function $p(x)$ by

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} L(X_i, x, \lambda), \tag{2.5}$$

and we estimate $g(x)$ by

$$\hat{g}(x) = \frac{n^{-1} \sum_{i=1}^{n} Y_i L(X_i, x, \lambda)}{\hat{p}(x)}. \tag{2.6}$$

When $\lambda_t = 0$ for all $t = 1, ..., r$, our estimator reverts back to the conventional approach whereby one uses a frequency estimator to deal with the discrete variables, while if $\lambda_t = 1$ for some $t$, then $\hat{g}(x)$ becomes unrelated to $x_t$. That is, $x_t$ is smoothed out from the regression model (it is deemed as an irrelevant variable).

We choose $\lambda = (\lambda_1, ..., \lambda_r)$ to minimize[2]

$$CV(\lambda) = \sum_{i=1}^{n} [Y_i - \hat{g}_{-i}(X_i)]^2, \tag{2.7}$$

where

$$\hat{g}_{-i}(X_i) = \frac{n^{-1} \sum_{j=1, j\neq i}^{n} Y_j L_{\lambda,ij}}{\hat{p}_{-i}(X_i)} \tag{2.8}$$

is the leave-one-out kernel estimator of $g(X_i)$, $L_{\lambda,ij} = L(X_i, X_j, \lambda)$, and

$$\hat{p}_{-i}(X_i) = \frac{1}{n} \sum_{j=1, j\neq i}^{n} L_{\lambda,ij} \tag{2.9}$$

is the leave-one-out estimator of $p(X_i)$. We will use $\hat{\lambda}$ to denote the cross-validation choice of $\lambda$ that minimizes (2.7).

We now present the assumptions that will be used to establish the asymptotic distribution of $\hat{g}(x)$.

**A1** (i) $\{X_i, Y_i\}_{i=1}^n$ are independent and identically distributed as $(X, Y)$. Let $\mathcal{D}$ denote the support of $X$, then $0 < \delta_1 \leq \min_{x\in\mathcal{D}} p(x) < \max_{x\in\mathcal{D}} p(x) \leq \delta_2 < \infty$, for some positive constants $\delta_1$ and $\delta_2$. Also $\max_{x\in\mathcal{D}} |g(x)| < \infty$. (ii) $E[Y_i^2 | X_i = x]$ is bounded on $x \in \mathcal{D}$.

**A2** Denote $d_{ij} = d_{x_i,x_j}$, $g_i = g(X_i)$, and define $B_1 = E\{E[(g_i - g_j)\mathbf{1}(d_{ij} = 1)|X_i]/p_i\}^2$, then $B_1 > 0$.

Assumption (A1) is quite standard. (A2) implies that $g(x)$ is not a constant function for $x \in \mathcal{D}$. It is needed prove that $\hat{\lambda} = o_p(1)$, and to establish the rate of convergence of $\hat{\lambda}$.

We first present a result when all regressors are relevant variables.

---

[2]For using least squares cross-validation to select smoothing parameters in a nonparametric regression model with continuous regressors, see Härdle and Marron (1985), and Härdle, Hall and Marron (1988, 1992).

**Theorem II.1** *Assume that $r_2 = 0$, that is all regressors are relevant variables, then under assumptions A1 and A2, we have*

$\hat{\lambda}_s = O_p(n^{-1})$ *for* $s = 1, ..., r_1$.

The proof of Theorem II.1 is much simpler than the proof of II.3 below, and therefore is omitted here. Theorem II.1 shows that, when all the regressors are relevant ones, the cross validation selected smoothing parameters converge to zero at a fast rate of $n^{-1}$. A proof of Theorem II.1 is available from the authors upon request. From II.1 one can easily obtain the following result.

**Theorem II.2** *Under the same conditions as in Theorem II.1, then*

$\sqrt{n}(\hat{g}(x) - g(x))/\sqrt{\hat{\Omega}(x)} \to N(0,1)$ *in distribution,*

*where* $\hat{\Omega}(x) = \hat{\sigma}^2(x)/\hat{p}(x)$, *and* $\hat{\sigma}^2(x) = n^{-1}\sum_i [Y_i - \hat{g}(X_i)]^2 L(X_i, x, \hat{\lambda})/\hat{p}(x)$ *is a consistent estimator of* $\sigma^2(x) = E[u_i^2 | X_i = x]$.

Proof: Define a frequency estimator $\tilde{g}(x)$ the same way as in $\hat{g}(x)$ but with $\lambda_t = 0$ for all $t = 1, ..., r_1$ (recall that $r_2 = 0$). Then it is well established that $\sqrt{n}(\tilde{g}(x) - g(x)) \to N(0, \Omega(x))$ in distribution, where $\Omega(x) = \sigma^2(x)/p(x)$. Next, using Theorem II.1, it is easy to see that $\hat{g}(x) = \tilde{g}(x) + O_p(n^{-1})$. Hence, $\sqrt{n}(\hat{g}(x) - g(x)) = \sqrt{n}(\tilde{g}(x) - g(x)) + O_p(n^{-1/2}) \to N(0, \Omega(x))$ in distribution. Finally, it is easy to show that $\hat{\Omega}(x) = \Omega(x) + o_p(1)$. Theorem II.2 follows.

**Theorem II.3** *Assume that $r_1 \geq 1$ and $r_2 \geq 1$ (with $r = r_1 + r_2 \geq 2$). Then under assumptions A1 and A2, we have*

$\hat{\lambda}_s = O_p(n^{-1/2})$ *for* $s = 1, ..., r_1$;

$lim_{n\to\infty} Pr\left(\hat{\lambda}_s = 1\right) = \alpha$ *for some* $\alpha \in (0,1)$ *for* $s = r_1 + 1, ..., r$.

The proof of Theorem II.3 is given in the subsection E.1.

Theorem II.3 states that the smoothing parameters associated with the relevant variables all converge to zero at the rate of $n^{-1/2}$, while the smoothing parameters for the irrelevant variables have a positive probability of taking the upper bound value of 1, that is, there is a positive probability that the irrelevant will be smoothed out. It is difficult to determine the exact value of $\alpha$ for the general case. However, when $u_i$ is symmetrically distributed around zero and is independent of $X_i$, then it can be show that $\alpha > 0.5$. Our simulation shows there is usually about a 60% chance that the $\hat{\lambda}_s$ takes the upper extreme value 1, with the remaining 40% chance that $\hat{\lambda}_s$ takes values between 0 and 1, for $s = r_1 + 1, ..., r$.

**Ordered Discrete Variables**

We know discuss the case that some of the discrete variables have an natural ordering. For an ordered discrete variable $X_{i,t}$ taking $c_t$ different values, Aitchison and Aitken (1976) suggest using $l(X_{i,t}, x_t, \lambda_t) = (1 - \lambda_t)$ if $X_{i,t} = x_t$, and $l(X_{i,t}, x_t, \lambda_t) = \lambda_t/(c_t - 1)$ if $X_{i,t} \neq x_t$. However, when $c_t \geq 3$, this kernel function has the problem that there does not exist a value of $\lambda_t$ such that $l(X_{i,t}, x_t, \lambda_t) =$ a constant. Hence, even when $x_t$ is an irrelevant variable, one cannot smooth out $x_t$. In this chapter we suggest a simple kernel weight function. For an ordered variable, we suggest using the following kernel:

$$\tilde{l}(\tilde{X}_{i,t}, \tilde{x}_t, \lambda_t) = \begin{cases} 1, & \text{if } \tilde{X}_{i,t} = \tilde{x}_t, \\ \lambda_t^{|\tilde{X}_{i,t} - \tilde{x}_t|}, & \text{if } \tilde{X}_{i,t} \neq \tilde{x}_t, \end{cases} \tag{2.10}$$

When $\lambda_t = 0$, we get an indicator function, and when $\lambda_t = 1$, we get a uniform weight function. Therefore, the range of $\lambda_t$ is $[0, 1]$. When $\lambda_t$ takes the upper bound value of 1, $x_t$ becomes an irrelevant variable (it is completely smoothed out). When some of the variables are ordered discrete variables, we use the kernel function defined

in (5.4), then it can be shown that the conclusion of Theorem II.3 remains unchanged. That is, $\hat{\lambda}_t = O_p(n^{-1/2})$ when $x_t$ is a relevant variable, and $\hat{\lambda}_t$ has a positive probability of taking the upper extreme value of 1 when $x_t$ is an irrelevant variable.

## C. Monte Carlo Simulations

We consider two data generating processes (DGP):

$$\begin{aligned}
\text{DGP}_1: &\quad Y_i = X_{i1} + X_{i2} + X_{i3} + X_{i1}X_{i2} + X_{i1}X_{i3} + X_{i2}X_{i3} + \epsilon_i, \\
\text{DGP}_2: &\quad Y_i = X_{i2} + X_{i3} + X_{i2}X_{i3} + \epsilon_i,
\end{aligned}$$

where $X_1, X_2, X_3 \in \{0, 1\}$, $Pr[X_j = 1] = 0.5, j = 1, 2, 3$, $\epsilon \sim N(0, 1)$.

Note that, in both DGP's, we have only 8 discrete 'cells'.

For each DGP we construct the proposed nonparametric estimator, the frequency estimator, and the following parametric models:

$$\begin{aligned}
\text{MODEL 1}: Y_i &= \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i1}X_{i2} \\
&\quad + \beta_5 X_{i1}X_{i3} + \beta_6 X_{i2}X_{i3} + u_i, \qquad (2.11) \\
\text{MODEL 2}: Y_i &= \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + u_i, \qquad (2.12)
\end{aligned}$$

Note that the parametric model 1 is a correct specification for both DGP's,[3] while the parametric model 2 is incorrect for both DGP's in that the interaction terms are missing in parametric model 2.

We let $n_1 = 100$, and for each Monte Carlo replication we compute the predicted mean square error (PMSE) on an *independent* sample drawn from the same DGP of size $n_2 = 1,000$. We conduct $5,000$ Monte Carlo replications, and report median, 5th

[3]Here we view an overspecified model as a correct model specification as it leads to consistent estimation of the conditional mean function.

and 95th percentiles of the PMSE

Table I. Summary of MSE for $DGP_1$: Both Variables Relevant

| Median, 5th, and 95th Percentiles of MSE | | |
|---|---|---|
| NP CV | NP FREQ | Param (M1) |
| 1.082 | 1.083 | 1.071 |
| [0.988, 1.176] | [0.987, 1.177] | [0.980, 1.160] |

Table II. Summary of MSE for $DGP_1$: Both Variables Relevant, Parametric Model Missing Interaction Terms

| Median, 5th, and 95th Percentiles of MSE | | |
|---|---|---|
| NP CV | NP FREQ | Param (M2) |
| 1.082 | 1.083 | 1.234 |
| [0.987, 1.173] | [0.987, 1.174] | [1.135, 1.322] |

Table III. Summary of MSE for $DGP_2$: One Variable Relevant

| Median, 5th, and 95th Percentiles of MSE | | |
|---|---|---|
| NP CV | NP FREQ | Param (M1) |
| 1.049 | 1.085 | 1.073 |
| [0.962, 1.129] | [0.989, 1.178] | [0.979, 1.160] |

Table I reports the estimation results based on DGP1 and based on correctly specified (parametric or nonparametric) models. We can see that all the three estimators perform well as they are all consistent estimators (since they all based on the correct specification).

From Table II gives the estimation result for DGP1 with the parametric model being the linear model (model 2). We observe that the parametric model 2 gives significantly larger PMSE's compared with the nonparametric models. This is because that the parametric model 2 is misspecified for DGP1 (missing the interaction terms).

Finally, Table III is based on DGP2 and all parametric estimation is based on the parametric model 2. Therefore, all estimation models are correctly specified (though they may be overspecified). From Table 3 we observe that the nonparametric cross-validation based method has a smaller PMSE than that obtained from the nonparametric frequency method. This is because for DGP2, $X_{i1}$ is an irrelevant variable, our CV-based estimator has a high probability of smoothing out the irrelevant variable, hence it leads to a more efficient (in finite samples) estimation result than the frequency estimator. It is interesting to observe that our nonparametric CV-based estimator also has a smaller PMSE than the correctly specified parametric estimator. This is certainly a finite application sample result since we know that asymptotically, a parametric estimator based on a correctly specified model has a fast rate of convergence than a nonparametric estimator, which ensures that asymptotically, a parametric estimator should have smaller PMSE. From Table 1 we know that for our experiment with $n = 100$, the two methods have similar PMSE's, when $X_{i1}$ becomes irrelevant, the parametric model PMSE does not improve, while the nonparametric CV estimator improves since it smoothes out the irrelevant regressor, hence, in finite sample applications, it is possible that our nonparametric estimator has a smaller PMSE than that of a parametric estimator based on a correctly specified model, especially when there exists some irrelevant variables.

## D. Applications

We consider two empirical applications which demonstrate the usefulness of the proposed approach relative to common parametric specifications which appear in the literature.

We shall focus on the out-of-sample predictive performance of both the proposed estimator and common parametric specifications which have been used to model each dataset. For each dataset we apply a random shuffle and then split the shuffled data into a training and evaluation set having $n_1$ and $n_2$ observations respectively. We then estimate each model on the training data and generate predictions using the explanatory variables in the evaluation dataset. We then compute the squared prediction error as the square of the difference between the predicted and actual values of the dependent variable in the evaluation dataset. To avoid the criticism that our results may reflect a non-representative split of the data, we randomly shuffle and split the shuffled data 100 times, and report median and mean values of squared prediction errors based on these 100 training and evaluation samples.

### 1. Count Survival Data - Veteran's Administration Lung Cancer Trial

We consider the dataset taken from Kalbfleisch and Prentice (1980, pg 223-224) which models survival in days of cancer patients with six discrete explanatory variables being treatment type, cell type, Karnofsky score, months from diagnosis, age in years, and prior therapy. The dataset contains 137 observations, and the number of cells greatly exceed the number of observations. A detailed description for the data can be found in Kalbfleisch and Prentice (1980). Clearly, the conventional frequency nonparametric method cannot be used for this dataset. The goal here is to model the expected survival in days. The estimation sample is $n_1 = 132$, and the prediction sample is

$n_2 = 5$. We compute the out-of-sample $MSE = n_2^{-1} \sum_{i=1}^{n_2} (y_i - \hat{y}_i)^2$, where $\hat{y}_i$ is the predicted value of $E(y_i|x_i)$ computed using $x_i$ and $\{x_j, y_j\}_{j=1}^{n_1}$ (the independent estimation sample). We randomly split the sample into two sub-samples of size $n_1$ and $n_2$ 100 times. Table IV gives the out-of-sample squared prediction errors generated from the different estimation methods. As can be seen from Table IV, the average nonparametric squared prediction error is only 60% to 62% of those obtained from various parametric methods.

Table IV. Veteran's Lung Cancer Data

| Model | Median MSE | Mean MSE | $\frac{\text{Mean MSE}_{np}}{\text{Mean MSE}}$ |
|---|---|---|---|
| NONP | 9,770.5 | 17,614.9 | 100% |
| OLS | 13,831.7 | 28,422.6 | 62% |
| POISSON | 15,319.8 | 29,052.9 | 60% |

In Table V we report median values of the cross-validated smoothing parameters over the 100 sample splits outlined above. We report this table to underscore how cross-validation automatically removes variables in applied settings.

Table V. Summary of Median Smoothing Parameters [number of categories in brackets]

| $\lambda_1$ [2] | $\lambda_2$ [4] | $\lambda_3$ [12] | $\lambda_4$ [28] | $\lambda_5$ [40] | $\lambda_6$ [2] |
|---|---|---|---|---|---|
| 0.934 | 0.056 | 0.004 | 0.157 | 1.000 | 1.000 |

As can be seen, the 'relevant' explanatory variables for predicting survival are cancer cell-type ($\lambda_2$), Karnofsky score[4] ($\lambda_3$), and months from diagnosis ($\lambda_4$). The

[4]The Karnofsky score measures patient performance of activities of daily living.

remaining regressors are smoothed out by cross-validation.

### 2. Count Data - Fair's 1977 Extramarital Affairs Data

We consider Fair's (1978) dataset which models how often an individual engaged in extramarital affairs during the past year. This often cited paper with a complete description of the data is located at `http://fairmodel.econ.yale.edu/rayfair/pdf/1978A200.PDF`. The dataset contains 601 observations and has eight discrete explanatory variables given by sex, age, number of years married, have children or not, how religious, level of education, occupation, and how they rate their marriage. Following Greene (2000, pg 920-926), the parametric models we use include Probit, Tobit, and Poisson specifications. The estimation sample is $n_1 = 500$, and the prediction sample is $n_2 = 101$. The goal is to model the expected number of affairs. We split the sample into two sub-sample of size $n_1$ and $n_2$ to compute the out-of-sample squared prediction errors, and repeat this procedure of randomly splitting the sample 100 times. Table VI reports the median and mean out-of-sample squared prediction errors by different estimation methods. We observe that our nonparametric squared prediction error is 69% to 82% of those obtained by various parametric methods.

Table VI. Fair's 1977 Extramarital Affairs Data

| Model | Median MSE | Mean MSE | $\frac{\text{Mean MSE}_{np}}{\text{Mean MSE}}$ |
|---|---|---|---|
| NONP | 9.97 | 10.18 | 100% |
| OLS | 11.90 | 12.39 | 82% |
| PROBIT | 14.66 | 14.74 | 69% |
| TOBIT | 12.81 | 13.04 | 78% |
| POISSON | 12.88 | 12.99 | 78% |

In Table VII we report median values of the cross-validated smoothing parameters over the 100 sample splits outlined above.

Table VII. Summary of Median Smoothing Parameters [number of categories in brackets]

| $\lambda_1$ [2] | $\lambda_2$ [9] | $\lambda_3$ [8] | $\lambda_4$ [2] | $\lambda_5$ [5] | $\lambda_6$ [7] | $\lambda_7$ [7] | $\lambda_8$ [5] |
|---|---|---|---|---|---|---|---|
| 0.416 | 0.064 | 1.000 | 0.998 | 0.507 | 0.999 | 0.282 | 0.004 |

As can be seen, the most 'relevant' explanatory variables from a cross-validation perspective are age ($\lambda_2$) and how a person rates their marriage ($\lambda_8$), while 'irrrelevant' variables are number of years married ($\lambda_3$), having children or not ($\lambda_4$), and level of education ($\lambda_6$) appear to be 'irrelevant' for prediction of number of affairs, and the remaining variables appear to have some relevance again from a cross-validation perspective. Thus, the cross-validation smoothing parameters reveal useful information regarding the relative importance of the variables. By way of comparison, using a Tobit model, Fair obtained a t-statistic of 3.63 for the number of years married, while our nonparametric method removes this variable from the resulting estimate. The out-of-sample prediction results suggest that the Tobit model is likely to be misspecified.

The two examples reported above suggest that our data-driven nonparametric estimator can yield *better* out-of-sample predictions than commonly used parametric models even when the number of cells is large relative to the sample size, and clearly the conventional frequency nonparametric estimator cannot be used in such cases.

## E. Proofs of the Main Results

### 1. Proof of Theorem II.3

The nonparametric regression model is

$$y_i = g\left(\bar{x}_i\right) + u_i \tag{2.13}$$

Instead we estimate an overspecified nonparametric function $g\left(x\right)$ by

$$\hat{g}\left(x\right) = \frac{n^{-1}\sum_i Y_i L\left(x, X_i\right)}{\hat{p}\left(x\right)} \tag{2.14}$$

where $\hat{p}\left(x\right) = \frac{1}{n}\sum_i L\left(x, X_i\right), X_i = \left(\bar{X}_i, \tilde{X}_i\right)$. Under the condition of (2.2), we have $p(\bar{X}_i, \tilde{X}_i) = \bar{p}(\bar{X}_i)\tilde{p}(\tilde{X}_i)$. Denotes by $g_i = g(\bar{X}_i)$, $\hat{g}_{-i} = \hat{g}_{-i}(X_i)$, we have

$$\begin{aligned} CV\left(\lambda\right) &= \frac{1}{n}\sum_{i=1}^{n}\left(g_i - \hat{g}_{-i}\right)^2 + \frac{2}{n}\sum_{i=1}^{n} u_i\left(g_i - \hat{g}_{-i}\right) + \frac{1}{n}\sum_{i=1}^{n} u_i^2 \\ &= I_1 + I_2 + I_3 \end{aligned} \tag{2.15}$$

where the third term $I_3$ is unrelated to $\lambda$.

First we consider $I_1$. Define $\hat{m}_{1,-i} = \frac{1}{n-1}\sum_{j\neq i}\left(g_i - g_j\right)L_{ji}$, $\hat{m}_{2,-i} = \frac{1}{n-1}\sum_{j\neq i} u_j L_{ji}$, and $\hat{p}_{-i} = \frac{1}{n-1}\sum_{j\neq i} L_{ji}$, where $L_{ij} = L(X_i, X_j, \lambda)$. Then we have

$$\begin{aligned} I_1 &= \frac{1}{n}\sum_{i=1}^{n}\left(g_i - \hat{g}_{-i}\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\hat{m}_{1,-i}^2/\hat{p}_{-i}^2 + \frac{1}{n}\sum_{i=1}^{n}\hat{m}_{2,-i}^2/\hat{p}_{-i}^2 - \frac{2}{n}\sum_{i=1}^{n}\hat{m}_{1,-i}\hat{m}_{2,-i}/\hat{p}_{-i}^2 \\ &\equiv S_1 + S_2 - 2S_3 \end{aligned} \tag{2.16}$$

where the definition of $S_j$ $(j = 1, 2, 3)$ should be apparent.

Define an indicator function $I_{X_l = X_i} = 1$ if $X_l = X_i$, and 0 otherwise, and denotes

by $\bar{\lambda} = \sum_{s=1}^{r_1} \lambda_s$, then by Lemma II.1, Lemma II.2, Lemma II.3, we have

$$\begin{aligned} S_1 &= \sum_{x \in D} \left[ \sum_{s=1}^{r_1} \lambda_s \left( \sum_{\bar{x}_j : |\bar{x}_j - \bar{x}\ |_s = 1} \bar{p}\left(\bar{x}_j\right) \left( g\left(\bar{x}\ \right) - g\left(\bar{x}_j\right) \right) \right) \right]^2 \tilde{p}\ \ (\tilde{x}) \left[ \bar{p}\left(\bar{x}\right) \right]^{-1} \\ &\quad + o_p\left( |\bar{\lambda}|^2 \right), \end{aligned} \tag{2.17}$$

$$\begin{aligned} S_2 &= \frac{E\left(u_j^2\right)}{n-1} E\left(1/\bar{p}_i^2\right) E\left( \tilde{L}_{ij}^2 / \tilde{\mu}_{p,-i}^2 \right) \\ &\quad + \frac{2}{n\left(n-1\right)} \sum_i \sum_{j>i} u_i u_j E_l \left\{ I_{\bar{X}_l = \bar{X}_i} I_{\bar{X}_l = \bar{X}_j} \frac{1}{\bar{p}_i^2} \right\} \theta_{ij}\left(\tilde{\lambda}\right) \\ &\quad + o_p\left( n^{-(1+\delta)} \right) + O\left( \frac{1}{n} |\bar{\lambda}|^2 \right) + O\left( \frac{1}{n} |\bar{\lambda}| \right). \end{aligned} \tag{2.18}$$

where $\theta_{ij}\left(\tilde{\lambda}\right) = E_l \frac{\tilde{L}_{li} \tilde{L}_{lj}}{\left\{ E_l\left[ \tilde{L}\left(\tilde{X}_l,\ \tilde{X}_i\right) \right] \right\}^2}$, $E_l(.)$ denotes expectation with respect to $X_l$.

$$\begin{aligned} S_3 &= \sum_{s=1}^{r_1} \lambda_s \{ \frac{1}{n\left(n-1\right)\left(n-1\right)} \sum_{i \neq j \neq l} u_i \left( g_l - g_i \right) I_{\left\| \bar{X}_i - \bar{X}_l \right\|_s = 1} \\ &\quad \frac{1}{\bar{p}_i^2} I_{\left\| \bar{X}_j - \bar{X}_i \right\| = 0} \cdot \tilde{L}_{ji} \tilde{L}_{jl} / \tilde{\mu}_{p,-j}^2 \} + O\left( n^{-1/2} \bar{\lambda}^2 \right) + o_p\left( n^{-3/4} \bar{\lambda} \right) \end{aligned} \tag{2.19}$$

Combining (2.17), (2.18) and (2.19), we obtain

$$\begin{aligned} I_1 &= \frac{1}{n}\sum_{i=1}^{n}\left(g_i-\hat{g}_{-i}\right)^2 \\ &= \sum_{x\in D}\left[\sum_{s=1}^{r_1}\lambda_s\left(\sum_{x_j:|\bar{x}_j-\bar{x}|_s=1}p\left(x_j\right)\left(g\left(\bar{x}\right)-g\left(\bar{x}_j\right)\right)\right)\right]^2\tilde{p}\left(\tilde{x}\right)\ \left[\bar{p}\left(\bar{x}\right)\right]^{-1} \\ &\quad +\frac{E\left(u_j^2\right)}{n-1}E\left(1/\bar{p}_i^2\right)E\left(\tilde{L}_{ij}^2/\tilde{\mu}_{p,-i}^2\right) \\ &\quad +\frac{2}{n\left(n-1\right)}\sum_i\sum_{j>i}u_iu_jE_l\left\{I_{\bar{X}_l=\bar{X}_i}I_{\bar{X}_l=\bar{X}_j}\frac{1}{\bar{p}_i^2}\right\}\theta_{ij}\left(\tilde{\lambda}\right) \\ &\quad -\sum_{s=1}^{r_1}\bar{\lambda}_s\{\frac{2}{n\left(n-1\right)\left(n-1\right)}\sum_{i\neq j\neq l}u_i\left(g_l-g_i\right)I_{\left\|\bar{X}_l-\bar{X}_i\right\|_s=1} \\ &\quad \frac{1}{\bar{p}_i^2}I_{\left\|\bar{X}_j-\bar{X}_i\right\|=0}\cdot\tilde{L}_{ji}\tilde{L}_{jl}/\tilde{\mu}_{p,-j}^2\}+o_p\left(n^{-3/4}\bar{\lambda}\right) \\ &\quad +o_p\left(n^{-(1+\delta)}\right)+o_p\left(\bar{\lambda}^2\right) \end{aligned} \tag{2.20}$$

By Lemma II.4,

$$\begin{aligned} I_2 &= \sum_{s=1}^{r_1}\bar{\lambda}_s\frac{2}{n\left(n-1\right)}\sum_{i=1}^{n}\sum_{l\neq i}u_i\left(g_i-g_l\right)/\bar{p}_iI_{\left\|\bar{X}_l-\bar{X}_i\right\|_s=1}\frac{\tilde{L}_{il}}{E_1\left[\tilde{L}\left(\tilde{X}_1,\tilde{X}_i\right)\right]} \\ &\quad -\frac{4}{n\left(n-1\right)}\sum_{i=1}^{n}\sum_{l>i}u_iu_lI_{\left\|\bar{X}_l-\bar{X}_i\right\|=0}\tilde{L}_{il}/\left(\bar{p}_iE_1\left[\tilde{L}\left(\tilde{X}_1,\tilde{X}_i\right)\right]\right) \\ &\quad +O_p\left(n^{-1}\bar{\lambda}\right)+O_p\left(n^{-1/2}\bar{\lambda}^2\right)+O_p\left(n^{-3/2}\right) \end{aligned} \tag{2.21}$$

Combining (2.15), (2.20) and (2.21), we have,

$$
\begin{aligned}
CV(\lambda) &= \sum_{x\in D}\left[\sum_{s=1}^{r}\lambda_s\left(\sum_{\bar{x}_j:|\bar{x}_j-\bar{x}|_s=1}\bar{p}(\bar{x}_j)\left(g(\bar{x})-g(\bar{x}_j)\right)\right)\right]^2\tilde{p}(\tilde{x})\left[\bar{p}(\bar{x})\right]^{-1}\\
&+\frac{E\left(u_j^2\right)}{n-1}E\left(1/\bar{p}_i^2\right)E\left(\tilde{L}_{ij}^2/\tilde{\mu}_{p,-i}^2\right)\\
&+\frac{2}{n(n-1)}\sum_i\sum_{j>i}u_iu_jE_l\left\{I_{\bar{X}_l=\bar{X}_i}I_{\bar{X}_l=\bar{X}_j}\frac{1}{\bar{p}_i^2}\right\}\theta_{ij}\left(\tilde{\lambda}\right)\\
&-\sum_{s=1}^{r_1}\bar{\lambda}_s\{\frac{2}{n(n-1)(n-1)}\sum\sum\sum_{i\neq j\neq l}u_i(g_l-g_i)I_{\left\|\bar{X}_l-\bar{X}_i\right\|_s=1}\\
&\frac{1}{\bar{p}_i^2}I_{\|\bar{x}_j-\bar{x}_i\|=0}\cdot\tilde{L}_{ji}\tilde{L}_{jl}/\tilde{\mu}_{p,-j}^2\}\\
&+\sum_{s=1}^{r_1}\bar{\lambda}_s\{\frac{2}{n(n-1)}\sum_i\sum_{l\neq i}u_i(g_i-g_l)/\bar{p}_i\\
&I_{\left\|\bar{X}_l-\bar{X}_i\right\|_s=1}\cdot\frac{\tilde{L}_{il}}{E_1\left[\tilde{L}\left(\tilde{X}_1,\tilde{X}_i\right)\right]}\}\\
&-\frac{4}{n(n-1)}\sum_i\sum_{l>i}u_iu_lI_{\|\bar{X}_l-\bar{X}_i\|=0}\tilde{L}_{il}/\left(\bar{p}_iE_1\left[\tilde{L}\left(\tilde{X}_1,\tilde{X}_i\right)\right]\right)\\
&+o_p\left(n^{-(1+\delta)}\right)+O\left(\sqrt{\frac{1}{n}}\bar{\lambda}^2\right)+o_p\left(n^{-3/4}\bar{\lambda}\right)\\
&+\text{terms unrelated to }\lambda, \qquad (2.22)
\end{aligned}
$$

Taking derivative of $CV(\lambda)$ with respect to $\lambda_t$ gives

$$
\begin{aligned}
&\frac{\partial}{\partial\bar{\lambda}_t}CV(\lambda)=2\sum_{x\in D}\sum_{s=1}^{r_1}\bar{\lambda}_s\sum_{x_j:|\bar{x}_j-\bar{x}|_s=1}p(x_j)\left(g(\bar{x})-g_j\right)\tilde{p}(\tilde{x})\left[\bar{p}(\bar{x})\right]^{-1}\\
&\cdot\sum_{x_j:|\bar{x}_j-\bar{x}|_t=1}p(x_j)\left(g(\bar{x})-g_j\right)-\frac{2}{n(n-1)(n-1)}\sum_{i\neq j\neq l}u_i(g_l-g_i)\\
&\cdot I_{\left\|\bar{X}_l-\ \bar{X}_i\right\|_t=1}\frac{1}{\bar{p}_i^2}I_{\|\bar{X}_j-\bar{X}_i\|=0}\cdot\tilde{L}_{ji}\tilde{L}_{jl}/\tilde{\mu}_{p,-j}^2\\
&+\frac{2}{n(n-1)}\sum_i\sum_{l\ \neq i}u_i(g_i-g_l)/\bar{p}_iI_{\left\|\bar{X}_l-\bar{X}_i\right\|_t=1}\cdot\frac{\tilde{L}_{il}}{E_1\left[\tilde{L}\left(\tilde{X}_1,\ \tilde{X}_i\right)\right]}\\
&+o_p\left(n^{-1/2}\right) \qquad (2.23)
\end{aligned}
$$

Let $\frac{\partial}{\partial\bar{\lambda}_t}CV\left(\lambda\right)=0$, we get

$$\bar{\lambda}_s=O_p\left(n^{-1/2}\right)\text{ for }s=1,...,r_1. \tag{2.24}$$

Using (2.24), we have for $t=r_1+1,...,r$,

$$\begin{aligned}
&\frac{\partial}{\partial\tilde{\lambda}_t}CV\left(\lambda\right)\\
=\;&\frac{E\left(u_j^2\right)}{n-1}E\left(1/\bar{p}_i^2\right)\cdot\frac{\partial}{\partial\tilde{\lambda}_t}E\left(\tilde{L}_{ij}^2/\tilde{\mu}_{p,-i}^2\right)\\
+\;&\frac{2}{n\left(n-1\right)}\sum_i\sum_{j>i}u_iu_jE_l\left\{I_{\bar{X}_l=\bar{X}_i}I_{\bar{X}_l=\bar{X}_j}\frac{1}{\bar{p}_i^2}\right\}\cdot\frac{\partial}{\partial\tilde{\lambda}_t}\theta_{ij}\left(\tilde{\lambda}\right)\\
-\;&\sum_{s=1}^{r_1}\bar{\lambda}_s\{\frac{2}{n\left(n-1\right)\left(n-1\right)}\sum_{i\neq j\neq l}u_i\left(g_l-g_i\right)I_{\left\|\bar{X}_l-\bar{X}_i\right\|_s=1}\\
&\frac{1}{\bar{p}_i^2}I_{\left\|\bar{X}_j-\bar{X}_i\right\|=0}\cdot\frac{\partial}{\partial\tilde{\lambda}_t}\tilde{L}_{ji}\tilde{L}_{jl}/\tilde{\mu}_{p,-j}^2\}\\
+\;&\sum_{s=1}^{r_1}\bar{\lambda}_s\frac{2}{n\left(n-1\right)}\sum_i\sum_{l\neq i}u_i\left(g_i-g_l\right)/\bar{p}_iI_{\left\|\bar{X}_l-\bar{X}_i\right\|_s=1}\frac{\partial}{\partial\tilde{\lambda}_t}\frac{\tilde{L}_{il}}{E_1(\tilde{L}_{1,i})}\\
-\;&\frac{4}{n\left(n-1\right)}\sum_i\sum_{l>i}u_iu_l\frac{\partial}{\partial\tilde{\lambda}_s}I_{\left\|\bar{X}_l-\bar{X}_i\right\|=0}\tilde{L}_{il}/\left(\bar{p}_iE_1(\tilde{L}_{1,i})\right)\\
&+o\left(n^{-1}\right).
\end{aligned} \tag{2.25}$$

Each of the main terms on the right of (2.25) has an order of $n^{-1}$. Since $\frac{\partial}{\partial\tilde{\lambda}_t}E\left(\tilde{L}_{ij}^2/\tilde{\mu}_{p,-i}^2\right)<0$, and all the other major terms are zero mean $O_p(n^{-1})$ random variables (note that $\bar{\lambda}_s=O_p(n^{-1/2})$, hence

$$\lim_{n\to\infty}Pr\left(\frac{\partial}{\partial\tilde{\lambda}_t}CV\left(\lambda\right)<0\right)>\alpha \tag{2.26}$$

for some $\alpha\in(0,1)$. Therefore,

$$\lim_{n\to\infty}Pr\left(\bar{\lambda}_s=1\right)>\alpha. \tag{2.27}$$

Since the first term on the right of (2.25) is minimized at $\lambda_s=1$ for $s=r_1+1,...,r$

(it is negative), and all other terms may have zero mean, this suggests that it is likely that $\alpha \in (1/2, 1)$. Indeed our simulations show that $\alpha$ is around to 0.6 for a variety of data generating processes.

## 2. Some Useful Lemmas

Define $\mu_{p,i} \equiv E\left(\hat{p}_{-i}|X_i\right)$. Hence we have

$$\begin{aligned}
\mu_{p,i} &\equiv E\left(\hat{p}_{-i}|X_i\right) = E\left(\frac{1}{n-1}\sum_{j\neq i} L\left(X_j, X_i\right)\right) = E\left[L\left(X_1, X_i\right)|X_i\right] \\
&= E\left[\bar{L}\left(\bar{X}_1, \bar{X}_i\right)\right] E\left[\tilde{L}\left(\tilde{X}_1, \tilde{X}_i\right)\right] \\
&= E[I_{\left\|\bar{X}_1-\bar{X}_i\right\|=0} + \sum_{s=1}^{r_1} \lambda_s I_{\left\|\bar{X}_1-\bar{X}_i\right\|_s=1} \\
&\quad + \sum_{s,t} \lambda_s\lambda_t I_{\left\|\bar{X}_1-\bar{X}_i\right\|_{s,t}=1}]E\left[\tilde{L}\left(\tilde{X}_1, \tilde{X}_i\right)\right] + O_p(\bar{\lambda}^3) \\
&= \left\{\bar{p}_i + \sum_{s=1}^{r_1} \bar{\lambda}_s \sum_{\bar{x}_1\in D, \|\bar{x}_i-\bar{x}_1\|_s=1} \bar{p}\left(\bar{x}_1\right)\right. \\
&\quad \left. + \sum_{s,t} \bar{\lambda}_s\bar{\lambda}_t \sum_{\bar{x}_1\in D, \|\bar{x}_i-\bar{x}_1\|_{s,t}=1} \bar{p}\left(\bar{x}_1\right)\right\} E[\tilde{L}\left(\tilde{X}_1, \tilde{X}_i\right)] + O_p(\bar{\lambda}^3) \qquad (2.28)
\end{aligned}$$

where $\left\|\bar{X}_1 - \bar{X}_i\right\| = 0$ means that $\bar{X}_{it} = \bar{X}_{1t}$ for all $t = 1, ..., r$ ; $\left\|\bar{X}_1 - \bar{X}_i\right\|_s = 1$ means that $\begin{cases} \bar{X}_{is} \neq \bar{X}_{1s} \\ \bar{X}_{it} = \bar{X}_{1t} \quad \text{for all } t \neq s \end{cases}$

$\left\|\bar{X}_1 - \bar{X}_i\right\|_{s,t} = 1$ means that $\begin{cases} \bar{X}_{iv} \neq \bar{X}_{1v} \quad \text{for } v = t \text{ or } v = s \\ \bar{X}_{iv} = \bar{X}_{1v} \quad \text{otherwise} \end{cases}$

$$\begin{aligned}\frac{1}{\mu_{p,i}^2} &= \frac{1}{\left\{E\left[\tilde{L}\left(\tilde{X}_1,\tilde{X}_i\right)\right]\right\}^2}\cdot\frac{1}{\bar{p}_i^2\left(1+\frac{E\left[\bar{L}\left(\bar{X}_1,\ \bar{X}_i\right)\right]-\bar{p}_i}{\bar{p}_i}\right)^2} \\ &= \frac{1}{\left\{E\left[\tilde{L}\left(\tilde{X}_1,\ \tilde{X}_i\right)\right]\right\}^2\bar{p}_i^2} \\ &\quad\cdot\left\{1-2\frac{E\left[\bar{L}\left(\bar{X}_1,\bar{X}_i\right)\right]-\bar{p}_i}{\bar{p}_i}+O\left(\bar{\lambda}^2\right)\right\} \end{aligned} \tag{2.29}$$

where $\bar{\lambda}=\sum_{s=1}^{r_1}\bar{\lambda}_s$,and obviously the last term $O\left(\bar{\lambda}^2\right)$ is uniform in $\bar{\lambda}$ and $\bar{X}_1$

$\hat{p}_{-i}=\frac{1}{n-1}\sum_{j\neq i}L\left(X_j,X_i\right)$, and we have

$$\begin{aligned}\hat{p}_{-i}-\mu_{p,i} &= \frac{1}{n-1}\sum_{j\neq i}L\left(X_j,X_i\right)-\ \mu_{p,i} \\ &= O_p\left(n^{-1/2}\right)\ \text{uniformly in } \bar{\lambda} \text{ and } X_i \end{aligned} \tag{2.30}$$

$$\begin{aligned}\hat{p}_{-i} &= \mu_{p,i}+O_p\left(n^{-1/2}\right) \\ &= \left\{\bar{p}_i+\sum_{s=1}^{r_1}\bar{\lambda}_s\left[\sum_{\bar{x}_1\in D,\|\bar{x}_i-\bar{x}_1\|_s=1}\bar{p}\left(\bar{x}_1\right)\right]\right. \\ &\quad \left.+\sum_{s,t}\bar{\lambda}_s\bar{\lambda}_t\left[\sum_{\bar{x}_1\in D,\|\bar{x}_i-\bar{x}_1\|_{s,t}=1}\bar{p}\left(\bar{x}_1\right)\right]+...\right\}E\left[\tilde{L}\left(\tilde{X}_1,\tilde{X}_i\right)\right] \\ &\quad +O_p\left(n^{-1/2}\right) \end{aligned} \tag{2.31}$$

where the last equation holds since (2.28)

Let $p_i^* = \bar{p}_i \cdot E_1\left[\tilde{L}\ \left(\tilde{X}_1, \tilde{X}_i\right) | \tilde{X}_i\right]$

$$\begin{aligned}
& \hat{p}_{-i} - p_i^* \\
= \ & \hat{p}_{-i} - \bar{p}_i \cdot E_1\ \left[\tilde{L}\left(\tilde{X}_1, \tilde{X}_i\right) |\ \tilde{X}_i\right] \\
= \ & \left\{\sum_{s=1}^{r_1} \bar{\lambda}_s \left[\sum_{\bar{x}_1 \in D, \|\bar{x}_i - \bar{x}_1\|_s = 1} \bar{p}\left(\bar{x}_1\right)\right] + \sum_{s,t} \bar{\lambda}_s \bar{\lambda}_t \left[\sum_{\bar{x}_1 \in D, \|\bar{x}_i - \bar{x}_1\|_{s,t} = 1} \bar{p}\left(\bar{x}_1\right)\right]\right\} \\
& \cdot E_1\left[\tilde{L}\left(\tilde{X}_1, \tilde{X}_i\right) | \tilde{X}_i\right] + O_p\left(n^{-1/2}\right) + o_p\left(\bar{\lambda}^2\right) \qquad (2.32)
\end{aligned}$$

$$\begin{aligned}
& 1/p_i^* - 1/\hat{p}_{-i} \\
= \ & \frac{1}{p_i^*} - \frac{1}{p_i^* \ + \left(\hat{p}_{-i} - p_i^*\right)} = \frac{1}{p_i^*}\left(1 - \frac{1}{1 + \frac{\left(\hat{p}_{-i} - p_i^*\ \right)}{p_i^*}}\right) \\
= \ & \frac{1}{p_i^*}\left(\frac{\hat{p}_{-i} - p_i^*}{p_i\ ^*} - \left(\frac{\hat{p}_{-i} - p_i^*}{p_i^*}\right)^2 + ...\right) \\
= \ & \left\{\sum_{s=1}^{r_1} \bar{\lambda}_s \left[\sum_{\bar{x}_1 \in D, \left\|\bar{X}_i - \bar{x}_1\right\|_s = 1} \bar{p}\left(\bar{x}_1\right)\right] + \sum_{s,t} \bar{\lambda}_s \bar{\lambda}_t \left[\sum_{\bar{x}_1 \in D, \left\|\bar{X}_i - \bar{x}_1\right\|_{s,t} = 1} \bar{p}\left(\bar{x}_1\right)\right]\right\} \\
& \cdot E_1\left[\tilde{L}\left(\tilde{X}_1, \tilde{X}_i\right) | \tilde{X}_i\right] / \left(p_i^*\right)^2 + O_p\left(n^{-1/2}\right) + o_p\left(\bar{\lambda}^2\right) \qquad (2.33)
\end{aligned}$$

**Lemma II.1**

$$S_1 = \sum_{x \in D}\left[\sum_{s=1}^{r_1} \lambda_s \left(\sum_{\bar{x}_j : |\bar{x}_j - \bar{x}|_s = 1} \bar{p}\left(\bar{x}_j\right)\left(g\left(\bar{x}\right) - g\left(\bar{x}_j\right)\right)\right)\right]^2 \tilde{p}\left(\tilde{x}\right)\left[\bar{p}\left(\bar{x}\right)\right]^{-1} + o_p\left(\bar{\lambda}^2\right)$$

Proof: Define $S_1^0 \equiv \frac{1}{n}\sum_i m_{1,-i}^2/\mu_{p,i}^2$, then

$$\begin{aligned}
S_1^0 \ = \ & \frac{1}{n\left(n-1\right)^2}\sum_i \sum_{j \neq i}\left(g_i - g_j\right)^2 L_{ji}^2/\mu_{p,i}^2 \\
& + \frac{1}{n\left(n-1\right)\left(n-2\right)}\sum_i \sum_{j \neq i} \sum_{l \neq i, l \neq j}\left(g_i - g_j\right)\left(g_i - g_l\right) L_{ji} L_{li}/\mu_{p,i}^2 \\
\equiv \ & G_1 + G_2. \qquad (2.34)
\end{aligned}$$

$G_2 = \frac{1}{n(n-1)(n-2)} \sum_i \sum_{j \neq i} \sum_{l \neq i, l \neq j} (g_i - g_j)(g_i - g_l) L_{ji} L_{li} / \mu_{p,i}^2$ can be written as a third order U- statistic. Let $Q_{ijl}$ be the symmetrized version of

$(g_i - g_j)(g_i - g_l) L_{ji} L_{li} / \mu_{p,i}^2$ , $Q_{ij} \equiv E(Q_{ijl} | X_i, X_j)$ , $Q_i \equiv E(Q_{ijl} | X_i)$, then we have:

$$\begin{aligned} G_2 &= EQ_1 + \frac{3}{n} \sum_i (Q_i - EQ_1) + \frac{6}{n(n-1)} \sum_{i=1}^{n} \sum_{j>i} (Q_{ij} - Q_i - Q_j + EQ_1) \\ &\quad + \frac{6}{n(n-1)(n-2)} \sum_{i=1}^{n} \sum_{j>i} \sum_{l>j} (Q_{ijl} - Q_{ij} - Q_{jl} - Q_{li} \\ &\quad + Q_i + Q_j + Q_l - EQ_1) \\ &= J_0 + J_1 + J_2 + J_3 \end{aligned} \tag{2.35}$$

$$\begin{aligned} J_0 &= EQ_1 = E(Q_{ijk}) = E\left[ \left[ E\left( (g_i - g_j) L_{ji} | X_i \right) \right]^2 / \mu_{p,i}^2 \right] \\ &= E\left[ \left[ E\left( (g_i - g_j) \bar{L}\left( \bar{X}_j, \bar{X}_i \right) \tilde{L}\left( \tilde{X}_j, \tilde{X}_i \right) | X_i \right) \right]^2 / \mu_{p,i}^2 \right] \\ &= E[\left[ E\left( (g_i - g_j) \bar{L}\left( \bar{X}_j, \bar{X}_i \right) | X_i \right) \right]^2 \left[ E\tilde{L}\left( \tilde{X}_j, \tilde{X}_i \right) | X_i \right]^2 \\ &\quad \left\{ E\left[ \bar{L}\left( \bar{X}_j, \bar{X}_i \right) | X_i \right] E\left[ \tilde{L}\left( \tilde{X}_j, \tilde{X}_i \right) | X_i \right] \right\}^2 ] \\ &= E\left[ \left[ E\left( (g_i - g_j) \bar{L}\left( \bar{X}_j, \bar{X}_i \right) | X_i \right) \right]^2 / \left\{ E\left[ \bar{L}\left( \bar{X}_j, \bar{X}_i \right) | X_i \right] \right\}^2 \right] \end{aligned} \tag{2.36}$$

Using Taylor expansion, we have, uniformly in $\bar{\lambda}$ and $x \in \mathcal{D}$

$$\begin{aligned} E\left( (g_i - g_j) \bar{L}_{ji} | X_i = x \right) &= \sum_{s=1}^{r_1} \bar{\lambda}_s \left( \sum_{\bar{x}_j : |\bar{x}_j - \bar{x}|_s = 1} \bar{p}(\bar{x}_j) \left( g(\bar{x}) - g(\bar{x}_j) \right) \right) \\ &\quad + O\left( \bar{\lambda}^2 \right), \end{aligned} \tag{2.37}$$

where $\bar{\lambda} = \sum_{s=1}^{r_1} \bar{\lambda}_s$.

Hence,

$$\begin{aligned} J_0 &= \sum_{x\in D} p(x)\left[\sum_{s=1}^{r_1}\bar{\lambda}_s\left(\sum_{\bar{x}_j:|\bar{x}_j-\bar{x}|_s=1}\bar{p}(\bar{x}_j)\left(g(\bar{x})-g(\bar{x}_j)\right)\right)\right]^2 / \left\{E\left[\bar{L}(\bar{x}_j,\bar{x})\ \right]\right\}^2 \\ &\quad +o\left(\bar{\lambda}^2\right) \\ &= \sum_{x\in D}\left[\sum_{s=1}^{r_1}\bar{\lambda}_s\left(\sum_{\bar{x}_j:|\bar{x}_j-\bar{x}|_s=1}\bar{p}(\bar{x}_j)\left(g(\bar{x}\ )-g(\bar{x}_j)\right)\right)\right]^2 p(x)\left\{E\left[\bar{L}(\bar{x}_j,\bar{x}\ )\ \right]\right\}^{-2} \\ &\quad +o\left(\bar{\lambda}^2\right) \\ &= \sum_{x\in D}\left[\sum_{s=1}^{r_1}\bar{\lambda}_s\left(\sum_{\bar{x}_j:|\bar{x}_j-\bar{x}|_s=1}\bar{p}(\bar{x}_j)\left(g(\bar{x})-g(\bar{x}_j)\right)\right)\right]^2 \tilde{p}(\tilde{x})\left[\bar{p}(\bar{x})\right]^{-1} \\ &\quad +o_p\left(\bar{\lambda}^2\right) \end{aligned} \tag{2.38}$$

where the last equality holds since $E\left[\bar{L}\left(\bar{x}_j,\bar{X}\right)\right]=\bar{p}(\bar{x})+O\left(\bar{\lambda}\right)$ uniformly in $x\in\mathcal{D}$.

Next we consider $J_1$.

$Q_i\equiv E\left(Q_{ijl}|X_i\right)$ has the same order as $\bar{\lambda}^2$, since

$E\left[\left(g_i-g_j\right)\left(g_i-g_l\right)L_{ji}L_{li}/\mu_{p,i}^2|X_i\right]=E\left[\left(g_i-g_j\right)L_{ji}|X_i\right]\}^2/\mu_{p,i}^2=O\left(\bar{\lambda}^2\right)$.

Therefore, $E\left(Q_i^k\right)=O\left(\bar{\lambda}^{2k}\right)$, by Rosenthal's inequality,

$E\left|J_1\right|^{2k}\leq n^{-2k}C_k\left(n^k\bar{\lambda}^{4k}+n\bar{\lambda}^{4k}\right)=O\left(n^{-k}\bar{\lambda}^{4k}\right)$

$P\left(J_1>n^{-\delta}\bar{\lambda}^2\right)\leq n^{-(1-2\delta)k}$ for some $0<\delta<1/2$.

So, $P\left(\sup_\lambda\left|J_1\right|\bar{\lambda}^{-2}>n^{-\delta}\right)\leq n^{-c}$

$$J_1=o_p\left(\bar{\lambda}^2\right)\ \text{ uniformly in } \lambda \tag{2.39}$$

Next, we consider $J_2$. Note that $Q_{ij}$ has the same order as

$E\left[\left(g_i-g_j\right)\left(g_i-g_l\right)L_{ji}L_{li}w\left(x_i\right)/\mu_{p,i}^2|x_l\right]\sim O\left(\bar{\lambda}\right)\left(g_i-g_j\right)L_{ji}w\left(x_i\right)$

So, $E\left(Q_{ij}^k\right)\sim O\left(\bar{\lambda}^{2k}\right)$

Hence,straightforward calculation shows that

$$E\left|J_2\right|^{2k}\leq n^{-4k}\left\{C_{k,1}n^{2k}\bar{\lambda}^{4k}+C_{k,2}n^{2k-2}\bar{\lambda}^{4k}+...+\right\}\sim n^{-2k}\bar{\lambda}^{4k}$$

$$P\left(J_2 > n^{-\delta}\bar{\lambda}^2\right) \leq n^{-(2-2\delta)k}$$

for some $0 < \delta < 1$.

$$P\left(\sup_{\lambda} |J_2| \, \bar{\lambda}^{-2} > n^{-\delta}\right) \leq n^{-c}$$

$$J_2 = o_p\left(\bar{\lambda}^2\right) \text{ uniformly in } \lambda \tag{2.40}$$

Now, we consider $J_3$

$$\begin{aligned} E\left(Q_{ijl}^k\right) &= E\left[\left(g_i - g_j\right)\left(g_i - g_l\right) L_{ji} L_{li} / \mu_{p,i}^2\right]^k \\ &= E\left\{E\left[\left(g_i - g_j\right)\left(g_i - g_l\right) L_{ji} L_{li} / \mu_{p,i}^2\right]^k | X_i\right\} \\ &= E\left\{E\left(g_i - g_j\right)^{2k} L_{ji}^{2k} / \mu_{p,i}^2 | X_i\right\} \\ &= O\left(\bar{\lambda}^{2k}\right) \end{aligned}$$

Like in the $J_2$ case,straightforward calculation shows that

$$E\left|J_3^{2k}\right| \leq n^{-6k}\left\{C_{k,1} n^{3k} \bar{\lambda}^{4k} + C_{k,2} n^{3k-3} \bar{\lambda}^{4k} + ...\right\} \sim n^{-3k} \bar{\lambda}^{4k}$$

$$P\left(J_3 > n^{-\delta}\bar{\lambda}^2\right) \leq n^{-(3-2\delta)k}$$

for some $0 < \delta < 1$.

$$P\left(\sup J_3 \bar{\lambda}^{-2} > n^{-\delta}\right) \leq n^{-(3-2\delta)k},$$

for some $0 < \delta < 1$. Hence,

$$J_3 = o_p\left(\bar{\lambda}^2\right) \text{ uniformly in } \bar{\lambda}. \tag{2.41}$$

Combining (2.35), (2.38), (2.39), (2.40), (2.41), we get

$$G_2 = \sum_{x\in D}\left[\sum_{s=1}^{r_1}\bar{\lambda}_s\left(\sum_{\bar{x}_j:|\bar{x}_j-\bar{x}\ |_s=1}\bar{p}\left(\bar{x}_j\right)\left(g\left(\bar{x}\ \right)-g\left(\bar{x}_j\right)\right)\right)\right]^2\tilde{p}\ \left(\tilde{x}\right)\left[\bar{p}\left(\bar{x}\right)\right]^{-1} + o_p\left(\bar{\lambda}^2\right) \tag{2.42}$$

Now consider $G_1$

$$G_1 = \frac{1}{n\left(n-1\right)^2}\sum_{i=1}^{n}\sum_{j\neq i}\left(g_i-g_j\right)^2 L_{ji}^2/\mu_{p,i}^2$$

Define $\Omega_{ij} = \left(g_i-g_j\right)^2 L_{ji}^2\left[1/\mu_{p,i}^2+1/\mu_{p,j}^2\right]/2$, $\Omega_i = E\left(\Omega_{ij}|X_i\right)$. Then

$$\begin{aligned} G_1 &= \frac{1}{n-1}\left[E\Omega_1\ +\frac{2}{n}\sum_{i=1}^{n}\left(\Omega_i-E\Omega_1\right)\right. \\ &\quad \left.+\frac{2}{n\left(n-1\right)}\sum_{i=1}^{n}\sum_{j>i}\left(\Omega_{ij}-\Omega_i-\Omega_j+E\Omega_1\right)\right] \\ &= G_{1,0}+G_{1,1}+G_{1,2} \end{aligned} \tag{2.43}$$

It is easy to see that,

$$G_{1,0} = O(n^{-1}\bar{\lambda}^2) \text{ uniformly in } \lambda \tag{2.44}$$

and

$$E\left(\Omega_i^k\right) = O(\bar{\lambda}^{2k})$$

$$E\left|G_{1,1}^{2k}\right| \leq n^{-4k}\left(n^k\bar{\lambda}^{4k}+n\bar{\lambda}^{4k}\right) \sim n^{-3k}\bar{\lambda}^{4k}$$

$$P\left(|G_{1,1}| > n^{-\delta}\bar{\lambda}^2\right) \leq n^{-(3-2\delta)k}, 0<\delta<1$$

$$G_{1,1} = o_p\left(\bar{\lambda}^2\right) \text{ uniformly in } \lambda \tag{2.45}$$

Similarly, one can show that,

$$G_{1,2} = o_p\left(\bar{\lambda}^2\right) \text{ uniformly in } \lambda \tag{2.46}$$

So, by (2.43), (2.44), (2.45), (2.46) we get that,

$$G_1 = o_p\left(\bar{\lambda}^2\right) \tag{2.47}$$

From (2.34), (2.42), (2.47), we have

$$\begin{aligned} S_1^0 &= \sum_{x\in D}\left[\sum_{s=1}^{r_1}\lambda_s\left(\sum_{x_j:|\bar{x}_j-\bar{x}|_s=1} p\left(x_j\right)\left(g\left(\bar{x}\right)-g\left(\bar{x}_j\right)\right)\right)\right]^2 \tilde{p}\left(\tilde{x}\right)\left[\bar{p}\left(\bar{x}\right)\right]^{-1} \\ &\quad +o\left(\bar{\lambda}^2\right) \text{ uniformly in } \lambda \end{aligned} \tag{2.48}$$

$$\left|S_1 - S_1^0\right| = \left|\frac{1}{n}\sum_{i=1}^{n}\hat{m}_{1,-i}^2\left(\frac{1}{\hat{p}_{-i}^2}-\frac{1}{\mu_{p,i}^2}\right)\right| \leq C \sup_{1\leq i\leq n}\hat{m}_{1,-i}^2 \sup_{1\leq i\leq n}\left|\Delta_{p,-i}\left(X_i\right)\right| \tag{2.49}$$

where

$$\Delta_{p,-i}\left(X_i\right) \equiv \hat{p}_{-i} - \mu_{p,i}$$

We define $\hat{m}_{1,-i}\left(x\right) = \frac{1}{n-1}\sum_{j\neq i}\left(g\left(\bar{x}\right)-g_j\right)L_{X_j,x}$, $L_{X_j,x} = L(X_j, x, \lambda)$.

$$\begin{aligned} E\left(\hat{m}_{1,-i}\left(x\right)\right) &= E\left(\left(g\left(\bar{x}\right)-g_j\right)L_{X_j,x}\right) \\ &= O\left(\bar{\lambda}\right), \text{uniformly in } x\in D \text{ and } \lambda. \end{aligned} \tag{2.50}$$

$\hat{m}_{1,-i}\left(x\right) - E\left(\hat{m}_{1,-i}\left(x\right)\right) = \frac{1}{n-1}\sum_{j\neq i}\left[\left(g\left(\bar{x}\right)-g_j\right)L_{X_j,x} - E\left(\left(g\left(\bar{x}\right)-g_j\right)L_{X_j,x}\right)\right] \equiv \frac{1}{n-1}\sum_{j\neq i}\left[\Phi_j - E\left(\Phi_j\right)\right]$, where $\Phi_j \equiv \left(g\left(\bar{x}\right)-g_j\right)L_{X_j,x}$ $E\left(\Phi_j^k\right) \sim O\left(\bar{\lambda}^k\right)$.

By Rothenthal′s inequality, we have

$$E\left|\hat{m}_{1,-i}\left(x\right) - E\left(\hat{m}_{1,-i}\left(x\right)\right)\right|^{2k} \leq C_k n^{-2k}\left(n^k\bar{\lambda}^{2k} + n\bar{\lambda}^{2k}\right) \sim O\left(n^{-k}\bar{\lambda}^{2k}\right)$$

By Markov's inequality,

$$P\left(\left|\hat{m}_{1,-i}(x) - E\left(\hat{m}_{1,-i}(x)\right)\right| > n^{-\delta}\bar{\lambda}\right) \leq n^{-(1-2\delta)k}, \qquad 0 < \delta < 1/2.$$

Hence,

$$P\left(\sup_{\lambda,x}\left|\hat{m}_{1,-i}(x) - E\left(\hat{m}_{1,-i}(x)\right)\right|\bar{\lambda}^{-1} > n^{-\delta}\right) \leq n^{-(1-2\delta)k}, 0 < \delta < 1/2.$$

Then we get,

$$\sup_{\lambda,x}\left|\hat{m}_{1,-i}(x) - E\left(\hat{m}_{1,-i}(x)\right)\right| = o_p\left(\bar{\lambda}\right) \tag{2.51}$$

So, by (2.50), (2.51)

$$\sup_{\lambda,x}\hat{m}_{1,-i}(x) = O_p\left(\bar{\lambda}\right). \tag{2.52}$$

By noting that $\bar{\lambda} = O_p(n^{-1/2})$, we have

$$\sup_{i,\lambda,x}\left|\Delta_{p,-i}(x)\right| \sup_{i,\lambda,x}\left|\Delta_{p,-i}(x)\right| = O_p\left(n^{-1/2}\right) \tag{2.53}$$

And by (2.52), (2.53), we get,

$$\left|S_1 - S_1^0\right| \leq C \sup_{1\leq i\leq n}\hat{m}_{1,-i}^2 \sup_{1\leq i\leq n}\left|\Delta_{p,-i}(X_i)\right| = o_p\left(\bar{\lambda}^2\right) \tag{2.54}$$

By (2.49), (2.54), we get

$$\begin{aligned} S_1 \;=\; & \sum_{x\in D}\left[\sum_{s=1}^{r_1}\lambda_s\left(\sum_{x_j:|\bar{x}_j-\bar{x}|_s=1} p(x_j)\left(g(\bar{x}) - g(\bar{x}_j)\right)\right)\right]^2 \tilde{p}(\tilde{x})\left[\bar{p}(\bar{x})\right]^{-1} \\ & + o_p\left(\bar{\lambda}^2\right) \end{aligned} \tag{2.55}$$

**Lemma II.2**

$$\begin{aligned} S_2 &= \frac{E\left(u_j^2\right)}{n-1} E\left(1/\bar{p}_i^2\right) E\left(\tilde{L}_{ij}^2/\tilde{\mu}_{p,-i}^2\right) \\ &\quad -\frac{2E\left(u_j^2\right)}{n-1}\sum_{s=1}^{r_1}\bar{\lambda}_s E\left(\frac{\bar{L}_{ij}^2\left[\sum_{\bar{x}_1\in D,\left\|\bar{X}_i-\bar{x}_1\right\|_s=1}\bar{p}\left(\bar{x}_1\right)\right]}{\bar{p}_i^3}\right) E\left(\tilde{L}_{ij}^2/\tilde{\mu}_{p,-i}^2\right) \\ &\quad +\frac{2}{n\left(n-1\right)}\sum_i\sum_{j>i} u_i u_j E_l\left\{I_{\bar{X}_l=\bar{X}_i} I_{\bar{X}_l=\bar{X}_j}\frac{1}{\bar{p}_i^2}\right\}\theta_{ij}\left(\tilde{\lambda}\right) \\ &\quad -\sum_{s=1}^{r_1}\bar{\lambda}_s\cdot\frac{1}{n}\bar{\Delta}_3\left(s,\tilde{\lambda}\right)+\sum_{s=1}^{r_1}\bar{\lambda}_s\cdot\frac{1}{n}\bar{\Delta}\left(s,\tilde{\lambda}\right)+o_p\left(n^{-(1+\delta)}\right)+O\left(\frac{1}{n}\bar{\lambda}^2\right)] \end{aligned}$$

Proof: $S_2=\frac{1}{n}\sum_{i=1}^n \hat{m}_{2,-i}^2/\hat{p}_{-i}^2$. Define $S_2^0=\frac{1}{n}\sum_{i=1}^n \hat{m}_{2,-i}^2/\mu_{p,-i}^2$. Then

$$\begin{aligned} S_2^0 &= \left[n\left(n-1\right)^2\right]^{-1}\sum_i\sum_{j\neq i} u_j^2 L_{ij}^2/\mu_{p,-i}^2 \\ &\quad +\left[n\left(n-1\right)^2\right]^{-1}\sum_i\sum_{j\neq i}\sum_{l\neq j\neq i} u_j L_{ij} u_l L_{il}/\mu_{p,-i}^2 \\ &= D_1+D_2. \end{aligned} \tag{2.56}$$

We consider $D_2$ first here.

Let $\Theta_{ijl}$ denote the symmetrized version of $u_j u_l L_{ij} L_{il} w_i/\mu_{p,-i}^2$, and define $\Theta_{ij}=E\left(\Theta_{ijl}|Z_i,Z_j\right)$, $Z_i=(X_i,u_i)$, then, we have

$$\begin{aligned} D_2 &= \frac{6}{n\left(n-1\right)}\sum_i\sum_{j>i}\Theta_{ij}+\frac{6}{n\left(n-1\right)\left(n-2\right)}\sum_i\sum_{j>i}\sum_{l>j>i}\left(\Theta_{ijl}-\Theta_{ij}-\Theta_{jl}-\Theta_{li}\right) \\ &= D_{21}+D_{22}. \end{aligned} \tag{2.57}$$

$$
\begin{aligned}
\Theta_{ij} &= E\left(\Theta_{ijl}|Z_i, Z_j\right) \\
&= \frac{1}{3}u_i u_j E_l\left(L_{li}L_{lj}/\mu^2_{p,-l}\right) = \frac{1}{3}u_i u_j E_l\{\left[I_{\bar{X}_l=\bar{X}_i} + \sum_{s=1}^{r_1}\lambda_s I_{\left\|\bar{X}_l-\ \bar{X}_i\right\|_s=1} + \cdots\right]\cdot \\
&\quad \left[I_{\bar{X}_l=\bar{X}_j} + \sum_{s=1}^{r_1}\lambda_s I_{\left\|\bar{X}_l-\bar{X}_j\right\|_s=1} + \cdots\right]\tilde{L}_{li}\ \tilde{L}_{lj}/\mu^2_{p,-l}\} \\
&= \frac{1}{3}u_i u_j E_l\left\{I_{\bar{X}_l=\bar{X}_i}I_{\bar{X}_l=\bar{X}_j}\tilde{L}_{li}\ \tilde{L}_{lj}/\mu^2_{p,-l}\right\} \\
&\quad + \frac{1}{3}u_i u_j \sum_{s=1}^{r_1}\lambda_s E_l\{[I_{\bar{X}_l=\bar{X}_i}I_{\left\|\bar{X}_l-\bar{X}_j\right\|_s=1} \\
&\quad + I_{\bar{X}_l=\bar{X}_j}I_{\left\|\bar{X}_l-\bar{X}_i\right\|_s=1}]\tilde{L}_{li}\ \tilde{L}_{lj}/\mu^2_{p,-l}\} + \frac{1}{3}u_i u_j O\left(\bar{\lambda}^2\right) \\
&= \Gamma_1 + \Gamma_2 + \Gamma_3 \qquad (2.58)
\end{aligned}
$$

where $\Gamma_3 = \frac{1}{3}u_i u_j O\left(\bar{\lambda}^2\right)$ and the definitions of $\Gamma_1$ and $\Gamma_2$ are hopefully apparent.

$$
\begin{aligned}
\Gamma_1 &= \frac{1}{3}u_i u_j E_l \left\{ I_{\bar{X}_l=\bar{X}_i} I_{\bar{X}_l=\bar{X}_j} \tilde{L}_{li}\tilde{L}_{lj}/\mu_{p,-l}^2 \right\} \\
&= \frac{1}{3}u_i u_j E_l \left\{ I_{\bar{X}_l=\bar{X}_i} I_{\bar{X}_l=\bar{X}_j} \tilde{L}_{li}\ \tilde{L}_{lj} \frac{1}{\left\{ E\left[ \tilde{L}\left( \tilde{X}_l,\ \tilde{X}_i \right) \right] \right\}^2 \bar{p}_i^2} \right. \\
&\quad \left. \left\{ 1 - 2\frac{E\left[ \bar{L}\left( \bar{X}_l, \bar{X}_i \right) \right] - \bar{p}_i}{\bar{p}_i} + O\left( \bar{\lambda}^2 \right) \right\} \right\} \\
&= \frac{1}{3}u_i u_j E_l \left\{ I_{\bar{X}_l=\bar{X}_i} I_{\bar{X}_l=\bar{X}_j} \frac{1}{\bar{p}_i^2} \right. \\
&\quad \left. \left\{ 1 - 2\frac{\sum_{s=1}^{r_1} \lambda_s \left[ \sum_{\bar{x}_1 \in D, \left\| \bar{X}_i - \bar{x}_1 \right\|_s = 1} \bar{p}\left( \bar{x}_1 \right) \right]}{\bar{p}_i} + O\left( \bar{\lambda}^2 \right) \right\} \right\} \\
&\quad \cdot \quad E_l \frac{\tilde{L}_{li}\tilde{L}_{lj}}{\left\{ E\left[ \tilde{L}\left( \tilde{X}_l, \tilde{X}_i \right) \right] \right\}^2} \\
&= \frac{1}{3}u_i u_j E_l \left\{ I_{\bar{X}_l=\bar{X}_i}\ I_{\bar{X}_l=\bar{X}_j} \frac{1}{\bar{p}_i^2} \right. \\
&\quad \left. \left\{ 1 - 2\frac{\sum_{s=1}^{r_1} \lambda_s \left[ \sum_{\bar{x}_1 \in D, \left\| \bar{X}_i - \bar{x}_1 \right\|_s = 1} \bar{p}\left( \bar{x}_1 \right) \right]}{\bar{p}_i} + O\left( \bar{\lambda}^2 \right) \right\} \right\} \theta_{ij}\left( \tilde{\lambda} \right) \\
&= \frac{1}{3}u_i u_j E_l \left\{ I_{\bar{X}_l=\bar{X}_i} I_{\bar{X}_l=\bar{X}_j} \frac{1}{\bar{p}_i^2} \right\} \theta_{ij}\left( \tilde{\lambda} \right) \\
&\quad -\frac{2}{3}\sum_{s=1}^{r_1} \lambda_s u_i u_j E_l \left\{ I_{\bar{X}_l=\bar{X}_i} I_{\bar{X}_l=\bar{X}_j} \frac{1}{\bar{p}_i^2} \frac{\left[ \sum_{\bar{x}_1 \in D, \left\| \bar{X}_i - \bar{x}_1 \right\|_s = 1} \bar{p}\left( \bar{x}_1 \right) \right]}{\bar{p}_i} \right\} \theta_{ij}\left( \tilde{\lambda} \right) \\
&\quad -\frac{2}{3}u_i u_j O\left( \bar{\lambda}^2 \right) \theta_{ij}\left( \tilde{\lambda} \right) \qquad (2.59)
\end{aligned}
$$

where $\theta_{ij}\left( \tilde{\lambda} \right) = \frac{E_l[\tilde{L}_{li}\tilde{L}_{lj}]}{\left\{ E\left[ \tilde{L}\left( \tilde{X}_l,\ \tilde{X}_i \right) \right] \right\}^2}$.

$$
\begin{aligned}
\Gamma_2 &= \frac{1}{3}u_iu_j\sum_{s=1}^{r_1}\lambda_sE_l\{[I_{\bar{X}_l=\bar{X}_i}I_{\left\|\bar{X}_l-\bar{X}_j\right\|_s=1}+I_{\bar{X}_l=\bar{X}_j}I_{\left\|\bar{X}_l-\bar{X}_i\right\|_s=1}]\tilde{L}_{li}\ \tilde{L}_{lj}/\mu_{p,-l}^2\} \\
&= \frac{1}{3}u_iu_j\sum_{s=1}^{r_1}\lambda_sE_l\left\{\left[I_{\bar{X}_l=\bar{X}_i}I_{\left\|\bar{X}_l-\ \bar{X}_j\right\|_s=1}+I_{\bar{X}_l=\bar{X}_j}I_{\left\|\bar{X}_l-\bar{X}_i\right\|_s=1}\right]\tilde{L}_{li}\tilde{L}_{lj}\right\} \\
&\quad \cdot\frac{1}{\left\{E\left[\tilde{L}\left(\tilde{X}_l,\ \tilde{X}_i\right)\right]\right\}^2\bar{p}_i^2}\left\{1-2\ \frac{E\left[\bar{L}\left(\bar{X}_l,\bar{X}_i\right)\right]-\bar{p}_i}{\bar{p}_i}+O\left(\bar{\lambda}^2\right)\right\} \\
&= \frac{1}{3}u_iu_j\sum_{s=1}^{r_1}\lambda_sE_l\left\{\left[I_{\bar{X}_l=\bar{X}_i}I_{\left\|\bar{X}_l-\bar{X}_j\right\|_s=1}+I_{\bar{X}_l=\ \bar{X}_j}I_{\left\|\bar{X}_l-\bar{X}_i\right\|_s=1}\right]\right\} \\
&\quad \frac{1}{\bar{p}_i^2}\left\{1-2\frac{E\left[\bar{L}\left(\bar{X}_l,\bar{X}_i\right)\right]-\bar{p}_i}{\bar{p}_i}+O\left(\bar{\lambda}^2\right)\right\}\}E_l\left[\frac{\tilde{L}_{li}\tilde{L}_{lj}}{\left\{E\left[\tilde{L}\left(\tilde{X}_l,\tilde{X}_i\right)\right]\right\}^2}\right] \\
&= \frac{1}{3}u_iu_j\sum_{s=1}^{r_1}\lambda_sE_l\left[I_{\bar{X}_l=\bar{X}_i}I_{\left\|\bar{X}_l-\bar{X}_j\right\|_s=1}+I_{\bar{X}_l=\bar{X}_j}I_{\left\|\bar{X}_l-\bar{X}_i\right\|_s=1}\right] \\
&\quad \cdot\frac{1}{\bar{p}_i^2}\left\{1-2\frac{E\left[\bar{L}\left(\bar{X}_l,\bar{X}_i\right)\right]-\bar{p}_i}{\bar{p}_i}\ +O\left(\bar{\lambda}^2\right)\right\}\theta_{ij}\left(\tilde{\lambda}\right) \\
&= \frac{1}{3}u_iu_j\sum_{s=1}^{r_1}\lambda_sE_l\left[I_{\bar{X}_l=\bar{X}_i}I_{\left\|\bar{X}_l-\bar{X}_j\right\|_s=1}+I_{\bar{X}_l=\bar{X}_j}I_{\left\|\bar{X}_l-\bar{X}_i\right\|_s=1}\right]\frac{1}{\bar{p}_i^2}\theta_{ij}\left(\tilde{\lambda}\right) \\
&\quad +\frac{1}{3}u_iu_j\sum_{s=1}^{r_1}\lambda_sE_l\left[I_{\bar{X}_l=\bar{X}_i}I_{\left\|\bar{X}_l-\ \bar{X}_j\right\|_s=1}+I_{\bar{X}_l=\bar{X}_j}I_{\left\|\bar{X}_l-\bar{X}_i\right\|_s=1}\right] \\
&\quad \cdot\frac{1}{\bar{p}_i^2}\left\{-2\frac{E\left[\bar{L}\left(\bar{X}_l,\bar{X}_i\right)\right]-\bar{p}_i}{\bar{p}_i}+O\left(\bar{\lambda}^2\right)\right\}\theta_{ij}\left(\tilde{\lambda}\right) \\
&= \sum_{s=1}^{r_1}\lambda_s\frac{1}{3}u_iu_jE_l\left[I_{\bar{X}_l=\bar{X}_i}I_{\left\|\bar{X}_l-\ \bar{X}_j\right\|_s=1}+I_{\bar{X}_l=\bar{X}_j}I_{\left\|\bar{X}_l-\bar{X}_i\right\|_s=1}\right]\frac{1}{\bar{p}_i^2}\theta_{ij}\left(\tilde{\lambda}\right) \\
&\quad +\frac{1}{3}u_iu_jO\left(\bar{\lambda}^2\right)\theta_{ij}\left(\tilde{\lambda}\right) \qquad (2.60)
\end{aligned}
$$

Hence, by (2.58, (2.59) and (2.60), we have

$$
\begin{aligned}
\Theta_{ij} &= \Gamma_1 + \Gamma_2 + \Gamma_3 \\
&= \frac{1}{3}u_i u_j E_l \left\{ I_{\bar{X}_l=\bar{X}_i} I_{\bar{X}_l=\bar{X}_j} \frac{1}{\bar{p}_i^2} \right\} \theta_{ij}\left(\tilde{\lambda}\right) \\
&\quad - \sum_{s=1}^{r_1} \lambda_s \frac{2}{3} u_i u_j E_l \left\{ I_{\bar{X}_l=\bar{X}_i} I_{\bar{X}_l=\bar{X}_j} \frac{1}{\bar{p}_i^2} \frac{\left[\sum_{\bar{x}_1 \in D, \left\|\bar{X}_i - \bar{x}_1\right\|_s = 1} \bar{p}\left(\bar{x}_1\right)\right]}{\bar{p}_i} \right\} \theta_{ij}\left(\tilde{\lambda}\right) \\
&\quad + \sum_{s=1}^{r_1} \lambda_s \frac{1}{3} u_i u_j E_l \{ [ I_{\bar{X}_l=\bar{X}_i} I_{\left\|\bar{X}_l - \bar{X}_j\right\|_s = 1} + I_{\bar{X}_l=\bar{X}_j} I_{\left\|\bar{X}_l - \bar{X}_i\right\|_s = 1} ] \frac{1}{\bar{p}_i^2} \theta_{ij}\left(\tilde{\lambda}\right) \\
&\quad + \frac{1}{3} u_i u_j O\left(\bar{\lambda}^2\right) \theta_{ij}\left(\tilde{\lambda}\right) + \frac{1}{3} u_i u_j O\left(\bar{\lambda}^2\right) \\
&= \frac{1}{3}u_i u_j E_l \left\{ I_{\bar{X}_l=\bar{X}_i} I_{\bar{X}_l=\bar{X}_j} \frac{1}{\bar{p}_i^2} \right\} \theta_{ij}\left(\tilde{\lambda}\right) \\
&\quad - \sum_{s=1}^{r_1} \lambda_s \frac{2}{3} u_i u_j E_l \left\{ I_{\bar{X}_l=\bar{X}_i} I_{\bar{X}_l=\bar{X}_j} \frac{1}{\bar{p}_i^2} \frac{\left[\sum_{\bar{x}_1 \in D, \left\|\bar{X}_i - \bar{x}_1\right\|_s = 1} \bar{p}\left(\bar{x}_1\right)\right]}{\bar{p}_i} \right\} \theta_{ij}\left(\tilde{\lambda}\right) \\
&\quad + \sum_{s=1}^{r_1} \lambda_s \frac{1}{3} u_i u_j E_l \{ [ I_{\bar{X}_l=\bar{X}_i} I_{\left\|\bar{X}_l - \bar{X}_j\right\|_s = 1} \\
&\quad + I_{\bar{X}_l=\bar{X}_j} I_{\left\|\bar{X}_l - \bar{X}_i\right\|_s = 1} ] \frac{1}{\bar{p}_i^2} \theta_{ij}\left(\tilde{\lambda}\right) + \frac{1}{3} u_i u_j O\left(\bar{\lambda}^2\right)
\end{aligned} \tag{2.61}
$$

$$
\begin{aligned}
D_{21} &= \frac{6}{n(n-1)} \sum_i \sum_{j>i} \Theta_{ij} \\
&= \frac{2}{n(n-1)} \sum_i \sum_{j>i} u_i u_j E_l \left\{ I_{\bar{X}_l=\bar{X}_i} I_{\bar{X}_l=\bar{X}_j} \frac{1}{\bar{p}_i^2} \right\} \theta_{ij} \left( \tilde{\lambda} \right) \\
&\quad + \sum_{s=1}^{r_1} \lambda_s \frac{2}{n(n-1)} \sum_i \sum_{j>i} u_i u_j E_l \{ [ I_{\bar{X}_l=\bar{X}_i} I_{\left\| \bar{X}_l - \bar{X}_j \right\|_s = 1} \\
&\quad + I_{\bar{X}_l=\bar{X}_j} I_{\left\| \bar{X}_l - \bar{X}_i \right\|_s = 1} ] \frac{1}{\bar{p}_i^2} \} \theta_{ij} \left( \tilde{\lambda} \right) \\
&\quad + \frac{2}{n(n-1)} \sum_i \sum_{j>i} u_i u_j O\left( \bar{\lambda}^2 \right) \\
&= \frac{2}{n(n-1)} \sum_i \sum_{j>i} u_i u_j E_l \left\{ I_{\bar{X}_l=\bar{X}_i} I_{\bar{X}_l=\bar{X}_j} \frac{1}{\bar{p}_i^2} \right\} \theta_{ij} \left( \tilde{\lambda} \right) \\
&\quad - \sum_{s=1}^{r_1} \lambda_s \cdot \frac{1}{n} \bar{\Delta}_3 \left( s, \tilde{\lambda} \right) + \sum_{s=1}^{r_1} \lambda_s \cdot \frac{1}{n} \bar{\Delta} \left( s, \tilde{\lambda} \right) + O\left( \frac{1}{n} \bar{\lambda}^2 \right). \qquad (2.62)
\end{aligned}
$$

$E\left( |D_{22}|^{2k} \right) = n^{-6k} \sum_{(t_2,\ldots,t_{2k}):2t_2+3t_3+\ldots+2kt_{2k}=2k, t_s \in \{0,1\}.} E\left[ \prod_{s=2}^{2k} \left( n^3 \Theta_{i_s j_s l_s}^s \right)^{t_s} \right]$
$\leq n^{-6k} C n^{3k} E\left( \Theta_{i_s j_s l_s}^2 \right)$, where $C$ is a constant not related to $n, k, \bar{\lambda}$.

$$
P\left( |D_{22}| > n^{-(1+\delta)} \right) \leq n^{2k(1+\delta)} E\left( \Theta_{i_s j_s l_s}^2 \right)^{-2k} C n^{-3k} \left( 1 + k\bar{\lambda} \right) \sim n^{-(1-2\delta)k}, \ 0 < \delta < 1/2
$$

Hence,

$$
\sup_{\lambda} |D_{22}| = o_p\left( n^{-(1+\delta)} \right) \qquad (2.63)
$$

By (2.62) and (2.63), we have

$$\begin{aligned} D_2 &= D_{21} + D_{22} \\ &= \frac{2}{n(n-1)} \sum_i \sum_{j>i} u_i u_j E_l \left\{ I_{\bar{X}_l = \bar{X}_i} I_{\bar{X}_l = \bar{X}_j} \frac{1}{\bar{p}_i^2} \right\} \theta_{ij} \left( \tilde{\lambda} \right) \\ &\quad - \sum_{s=1}^{r_1} \lambda_s \cdot \frac{1}{n} \bar{\Delta}_{\ 3} \left( s, \tilde{\lambda} \right) + \sum_{s=1}^{r_1} \lambda_s \cdot \frac{1}{n} \bar{\Delta} \left( s, \tilde{\lambda} \right) + O\left( \frac{1}{n}\ \bar{\lambda}^2 \right) \\ &\quad + o_p\left( n^{-(1+\delta)} \right) \end{aligned} \tag{2.64}$$

Now we consider $D_1$.

$D_1 = \left[ n(n-1)^2 \right]^{-1} \sum_i \sum_{j \neq i} u_j^2 L_{ij}^2 / \mu_{p,-i}^2$,

Define $V_{ij} = 1/2 \left[ u_j^2 L_{ij}^2 / \mu_{p,-i}^2 + u_i^2 L_{ij}^2 / \mu_{p,-j}^2 \ \right], V_i = E\left( V_{ij} | X_i, u_i \right)$

then we have,

$$\begin{aligned} D_1 &= \frac{1}{n-1} \left\{ EV_1 + \frac{2}{n} \sum_i \left[ V_i - EV_1 \ \right] + \frac{2}{n(n-1)} \sum_i \sum_{j>i} \left[ V_{ij} - V_i - V_j + EV_1 \right] \right\} \\ &= B_0 + B_1 + B_2 \end{aligned} \tag{2.65}$$

Obviously,

$$
\begin{aligned}
B_0 &= \frac{1}{n-1}EV_1 = \frac{1}{n-1}E\left(u_j^2 L_{ij}^2/\mu_{p,-i}^2\right) = \frac{1}{n-1}E\left(u_j^2 \bar{L}_{ij}^2/\bar{\mu}_{\ p,-i}^{\ 2}\right) E\left(\tilde{L}_{ij}^2/\tilde{\mu}_{p,-i}^2\right) \\
&= \frac{E\left(u_j^2\right)}{n-1}E\left(\bar{L}_{ij}^2\frac{1}{\bar{p}_i^2}\left\{1-2\frac{E\left[\bar{L}\left(\bar{X}_1,\ \bar{X}_i\right)\right]-\bar{p}_i}{\bar{p}_i}+O\left(\bar{\lambda}^2\right)\right\}\right) E\left(\tilde{L}_{ij}^2/\tilde{\mu}_{\ p,-i}^{\ 2}\right) \\
&= \frac{E\left(u_j^2\right)}{n-1}E\left(/\bar{p}_i^2\right) E\left(\tilde{L}_{ij}^2/\tilde{\mu}_{p,-i}^2\right) + O\left(\frac{1}{n}\ \bar{\lambda}^2\right) E\left(\tilde{L}_{ij}^2/\tilde{\mu}_{\ p,-i}^{\ 2}\right) \\
&\quad -2\frac{E\left(u_j^2\right)}{n-1}E\left(\bar{L}_{ij}^2\frac{\sum_{s=1}^{r_1}\bar{\lambda}_s\left[\sum_{\bar{x}\ _1\in D,\left\|\bar{X}_i-\bar{x}_1\right\|_s=1}\bar{p}\ \left(\bar{x}_1\right)\right]}{\bar{p}_i^3}\right) E\left(\tilde{L}_{\ ij}^{\ 2}/\tilde{\mu}_{p,-i}^2\right) \\
&= \frac{E\left(u_j^2\right)}{n-1}E\left(1/\bar{p}_i^2\right) E\left(\tilde{L}_{ij}^2/\tilde{\mu}_{p,-i}^2\right) \\
&\quad -\frac{2E\left(u_j^2\right)}{n-1}\sum_{s=1}^{r_1}\bar{\lambda}_{\ s}E\left(\frac{\bar{L}_{ij}^2\left[\sum_{\bar{x}_1\in D,\left\|\bar{X}_i-\bar{x}_1\right\|_s=1}\bar{p}\left(\bar{x}_1\right)\right]}{\bar{p}_i^3}\right) E\left(\tilde{L}_{\ ij}^{\ 2}/\tilde{\mu}_{p,-i}^2\right) \\
&\quad +O\left(\frac{1}{n}\bar{\lambda}^2\right) E\left(\tilde{L}_{ij}^2/\ \tilde{\mu}_{p,-i}^2\right) \qquad (2.66)
\end{aligned}
$$

Obviously,

$$
B_1 = o_p\left(\frac{1}{n^{3/2}}\right) \text{ and } B_2 = o_p\left(\frac{1}{n^{5/2}}\right) \qquad (2.67)
$$

By (2.65) to (2.67), we have

$$
\begin{aligned}
D_1 &= \frac{E\left(u_j^2\right)}{n-1}E\left(1/\bar{p}_i^2\right) E\left(\tilde{L}_{ij}^2/\tilde{\mu}_{p,-i}^2\right) \\
&\quad -\frac{2E\left(u_j^2\right)}{n-1}\sum_{s=1}^{r_1}\bar{\lambda}_{\ s}E\left(\frac{\bar{L}_{ij}^2\left[\sum_{\bar{x}_1\in D,\left\|\bar{X}_i-\bar{x}_1\right\|_s=1}\bar{p}\left(\bar{x}_1\right)\right]}{\bar{p}_i^3}\right) E\left(\tilde{L}_{\ ij}^{\ 2}/\tilde{\mu}_{p,-i}^2\right) \\
&\quad +O\left(\frac{1}{n}\bar{\lambda}^2\right) E\left(\tilde{L}_{ij}^2/\ \tilde{\mu}_{p,-i}^2\right) + o_p\left(\frac{1}{n^{3/2}}\right) + o_p\left(\frac{1}{n^{5/2}}\right) \qquad (2.68)
\end{aligned}
$$

By (2.56), (2.64) and (2.68), we have

$$\begin{aligned} S_2^0 &= D_1 + D_2 = \frac{E\left(u_j^2\right)}{n-1} E\left(1/\bar{p}_i^2\right) E\left(\tilde{L}_{ij}^2/\tilde{\mu}_{p,-i}^2\right) \\ &\quad - \frac{2E\left(u_j^2\right)}{n-1} \sum_{s=1}^{r_1} \bar{\lambda}_s E\left(\frac{\bar{L}_{ij}^2 \left[\sum_{\bar{x}_1 \in D, \left\|\bar{X}_i - \bar{x}_1\right\|_s = 1} \bar{p}\left(\bar{x}_1\right)\right]}{\bar{p}_i^3}\right) E\left(\tilde{L}_{ij}^2/\tilde{\mu}_{p,-i}^2\right) \\ &\quad + \frac{2}{n\left(n-1\right)} \sum_i \sum_{j>i} u_i u_j E_l \left\{ I_{\bar{X}_l = \bar{X}_i} I_{\bar{X}_l = \bar{X}_j} \frac{1}{\bar{p}_i^2} \right\} \theta_{ij}\left(\tilde{\lambda}\right) \\ &\quad - \sum_{s=1}^{r_1} \bar{\lambda}_s \cdot \frac{1}{n} \bar{\Delta}_3\left(s, \tilde{\lambda}\right) + \sum_{s=1}^{r_1} \bar{\lambda}_s \cdot \frac{1}{n} \bar{\Delta}\left(s, \tilde{\lambda}\right) \\ &\quad + o_p\left(n^{-(1+\delta)}\right) + O\left(\frac{1}{n}\bar{\lambda}^2\right) \end{aligned} \tag{2.69}$$

$\left|S_2 - S_2^0\right| \leq C \left|\frac{1}{n} \sum_{i=1}^n \hat{m}_{2,-i}^2 w\left(x_i\right)\right| \sup \left|\Delta_{p,-i}\left(x\right)\right| = C \left|\frac{1}{n} \sum_{i=1}^n \hat{m}_{2,-i}^2 w\left(x_i\right)\right| \cdot O_p\left(\sqrt{\frac{1}{n}}\right)$, but

$\frac{1}{n} \sum_{i=1}^n \hat{m}_{2,-i}^2 w\left(x_i\right) \sim S_2^0$, so,

$$S_2 = S_2^0 + \left(S_2 - S_2^0\right) \sim S_2^0$$

$$\begin{aligned} S_2 &= \frac{E\left(u_j^2\right)}{n-1} E\left(1/\bar{p}_i^2\right) E\left(\tilde{L}_{ij}^2/\tilde{\mu}_{p,-i}^2\right) \\ &\quad - \frac{2E\left(u_j^2\right)}{n-1} \sum_{s=1}^{r_1} \bar{\lambda}_s E\left(\frac{\bar{L}_{ij}^2 \left[\sum_{\bar{x}_1 \in D, \left\|\bar{X}_i - \bar{x}_1\right\|_s = 1} \bar{p}\left(\bar{x}_1\right)\right]}{\bar{p}_i^3}\right) E\left(\tilde{L}_{ij}^2/\tilde{\mu}_{p,-i}^2\right) \\ &\quad + \frac{2}{n\left(n-1\right)} \sum_i \sum_{j>i} u_i u_j E_l \left\{ I_{\bar{X}_l = \bar{X}_i} I_{\bar{X}_l = \bar{X}_j} \frac{1}{\bar{p}_i^2} \right\} \theta_{ij}\left(\tilde{\lambda}\right) \\ &\quad - \sum_{s=1}^{r_1} \bar{\lambda}_s \cdot \frac{1}{n} \bar{\Delta}_3\left(s, \tilde{\lambda}\right) + \sum_{s=1}^{r_1} \bar{\lambda}_s \cdot \frac{1}{n} \bar{\Delta}\left(s, \tilde{\lambda}\right) \\ &\quad + o_p\left(n^{-(1+\delta)}\right) + O\left(\frac{1}{n}\bar{\lambda}^2\right) \end{aligned} \tag{2.70}$$

**Lemma II.3**

$$\begin{aligned} S_3 &= \sum_{s=1}^{r_1} \bar{\lambda}_s \{ \frac{1}{n\left(n-1\right)\left(n-1\right)} \sum_{i \neq j \neq l} u_i \left(g_l - g_i\right) I_{\left\| \bar{X}_l - \bar{X}_i \right\|_s = 1} \\ & \frac{1}{\bar{p}_i^2} I_{\left\| \bar{X}_j - \bar{X}_i \right\| = 0} \cdot \tilde{L}_{ji} \tilde{L}_{jl} / \tilde{\mu}_{p,-j}^2 \} + O\left( \sqrt{\frac{1}{n}} \bar{\lambda}^2 \right) + o_p\left( n^{-3/4} \bar{\lambda} \right) \end{aligned}$$

Proof: Note that $S_3 = \frac{1}{n} \sum_{i=1}^{n} \hat{m}_{1,-i} \hat{m}_{2,-i} / \hat{p}_{-i}^2$. Define $S_3^0 = \frac{1}{n} \sum_{i=1}^{n} \hat{m}_{1,-i} \hat{m}_{2,-i} / \mu_{p,-i}^2$. Hence,

$$\begin{aligned} S_3^0 &= \frac{1}{n\left(n-1\right)\left(n-1\right)} \sum_{i=1}^{n} \left[ \sum_{j \neq i} \left(g_i - g_j\right) L_{ji} \right] \left[ \sum_{j \neq i} u_j L_{ji}, \right] / \mu_{p,-i}^2 \\ &= \frac{1}{n\left(n-1\right)\left(n-1\right)} \sum_{i=1}^{n} \left[ \sum_{j \neq i} u_j \left(g_i - g_j\right) L_{ji}^2 \right] / \mu_{p,-i}^2 \\ &\quad + \frac{1}{n\left(n-1\right)\left(n-1\right)} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{l \neq j \neq i} \left[ \left(g_i - g_j\right) L_{ji} \right] \left[ u_l L_{li}, \right] / \mu_{p,-i}^2 \\ &= M_1 + M_2 \end{aligned} \tag{2.71}$$

Obviously,

$$\begin{aligned} M_1 &= \frac{1}{n\left(n-1\right)\left(n-1\right)} \sum_{i=1}^{n} \sum_{j \neq i} u_j \left(g_i - g_j\right) L_{ji}^2 w\left(x_i\right) / \mu \\ &= O\left( n^{-3/2} \bar{\lambda}^2 \right) \end{aligned} \tag{2.72}$$

and

$$\begin{aligned} M_2 &= \frac{1}{n\left(n-1\right)\left(n-1\right)} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{l \neq j \neq i} \left[ \left(g_j - g_i\right) L_{ji} \right] \left[ u_l L_{li}, \right] / \mu_{p,-i}^2 \\ &= \sum_{s=1}^{r_1} \lambda_s \frac{1}{n\left(n-1\right)\left(n-1\right)} \sum \sum \sum_{i \neq j \neq l} u_l \left(g_j - g_i\right) I_{\left\| \bar{X}_j - \bar{X}_i \right\|_s = 1} \\ & \frac{1}{\bar{p}_i^2} I_{\left\| \bar{X}_l - \bar{X}_i \right\| = 0} \cdot \tilde{L}_{ji} \tilde{L}_{il} / \tilde{\mu}_{p,-i}^2 + O\left( n^{-1/2} \bar{\lambda}^2 \right) \end{aligned} \tag{2.73}$$

By (2.71), (2.72) and (2.73), we have

$$\begin{aligned} S_3^0 &= \sum_{s=1}^{r_1} \lambda_s \frac{1}{n(n-1)(n-1)} \sum\sum\sum_{i\neq j\neq l} u_l (g_j - g_i) I_{\left\|\bar{X}_j - \bar{X}_i\right\|_s = 1} \\ &\quad \frac{1}{\bar{p}_i^2} I_{\left\|\bar{X}_l - \bar{X}_i\right\| = 0} \cdot \tilde{L}_{ji} \tilde{L}_{il} / \tilde{\mu}_{p,-i}^2 + O\left(n^{-1/2} \bar{\lambda}^2\right) \end{aligned} \tag{2.74}$$

We also have

$|S_3 - S_3^0| \leq C \sup |\Delta_{p,-i}| \cdot \sup |\hat{m}_{1,-i}| \cdot \sup |\hat{m}_{2,-i}| \sim O_p\left(\sqrt{\frac{1}{n}}\right) \cdot O_p\left(\bar{\lambda}\right) \cdot o_p\left(n^{-1/4}\right) \sim o_p\left(n^{-3/4}\bar{\lambda}\right)$.

Hence,

$$\begin{aligned} S_3 &= \sum_{s=1}^{r_1} \bar{\lambda}_s \{ \frac{1}{n(n-1)(n-1)} \sum_{i\neq j\neq l} u_i (g_l - g_i) I_{\left\|\bar{X}_l - \bar{X}_i\right\|_s = 1} \\ &\quad \frac{1}{\bar{p}_i^2} I_{\left\|\bar{X}_j - \bar{X}_i\right\| = 0} \cdot \tilde{L}_{ji} \tilde{L}_{jl} / \tilde{\mu}_{p,-j}^2 \} + O\left(n^{-1/2} \bar{\lambda}^2\right) + o_p\left(n^{-3/4}\bar{\lambda}\right) \end{aligned} \tag{2.75}$$

**Lemma II.4**

$$\begin{aligned} I_2 &= \sum_{s=1}^{r_1} \bar{\lambda}_s \left\{ \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{l\neq i} u_i (g_i - g_l) / \bar{p}_i I_{\left\|\bar{X}_l - \bar{X}_i\right\|_s = 1} \cdot \frac{\tilde{L}_{il}}{E_1\left[\tilde{L}\left(\tilde{X}_1, \tilde{X}_i\right)\right]} \right\} \\ &\quad - \frac{4}{n(n-1)} \sum_{i=1}^{n} \sum_{l>i} u_i u_l I_{\left\|\bar{X}_l - \bar{X}_i\right\| = 0} \tilde{L}_{il} / \left(\bar{p}_i E_1\left[\tilde{L}\left(\tilde{X}_1, \tilde{X}_i\right)\right]\right) \\ &\quad + O_p\left[\sqrt{\frac{1}{n}} \bar{\lambda}^2\right] + O_p\left(\frac{1}{n}\bar{\lambda}\right) + O_p\left(n^{-3/2}\right) \end{aligned}$$

Proof:

$$\begin{aligned} I_2 &= \frac{2}{n} \sum_{i=1}^{n} u_i (g_i - \hat{g}_{-i}(x_i)) \\ &= \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{l\neq i} u_i (g_i - g_l) L_{il} / \hat{p}_{-i} - \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{l\neq i} u_i u_l L_{il} / \hat{p}_{-i} \\ &= T_1 - T_2. \end{aligned} \tag{2.76}$$

Define $T_1^0 \equiv \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{l\neq i} u_i (g_i - g_l) L_{il} / \mu_{p,-i}$,

$T_1^1 \equiv \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{l\neq i} u_i (g_i - g_l) L_{il} / p_i$,

$T_2^1 \equiv \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{l \neq i} u_i u_l L_{il} / \ p_i^*,$

$$
\begin{aligned}
T_2^1 &= \frac{4}{n\left(n-1\right)} \sum_i \sum_{l>i} u_i u_l L_{il} / \ p_i^* \\
&= \frac{4}{n\left(n-1\right)} \sum_i \sum_{l>i} u_i u_l [I_{\left\| \bar{X}_l - \bar{X}_i \right\| = 0} + \sum_{s=1}^{r_1} \bar{\lambda}_s I_{\left\| \bar{X}_l - \bar{X}_i \right\|_s = 1} \\
&\quad + \sum_{s,t} \bar{\lambda}_s \bar{\lambda}_t I_{\left\| \bar{X}_l - \bar{X}_i \right\|_{s,t} = 1} + ...] \tilde{L}_{il} / \left( \bar{p}_i E_1 \left[ \tilde{L} \left( \tilde{X}_1, \tilde{X}_i \right) \right] \right) \\
&= \frac{4}{n\left(n-1\right)} \sum_i \sum_{l>i} u_i u_l I_{\left\| \bar{X}_l - \bar{X}_i \right\| = 0} \tilde{L}_{il} / \left( \bar{p}_i E_1 \left[ \tilde{L} \left( \tilde{X}_1, \tilde{X}_i \right) \right] \right) \\
&\quad + \sum_{s=1}^{r_1} \bar{\lambda}_s \left\{ \frac{4}{n\left(n-1\right)} \sum_i \sum_{l>i} u_i u_l \cdot I_{\left\| \bar{X}_l - \bar{X}_i \right\|_s = 1} \cdot \tilde{L}_{il} / \left( \bar{p}_i E_1 \left[ \tilde{L} \left( \tilde{X}_1, \tilde{X}_i \right) \right] \right) \right\} \\
&\quad + O\left( \frac{1}{n} \bar{\lambda}^2 \right) \qquad (2.77)
\end{aligned}
$$

$$
\begin{aligned}
T_2 - T_2^1 &= \frac{2}{n\left(n-1\right)} \sum_i \sum_{l \neq i} u_i u_l L_{il} / \hat{p}_{-i} - \frac{2}{n\left(n-1\right)} \sum_i \sum_{l \neq i} u_i u_l L_{il} / p_i^* \\
&= -\frac{2}{n\left(n-1\right)} \sum_i \sum_{l \neq i} u_i u_l L_{il} \frac{1}{p_i^*} \left( \frac{\hat{p}_{-i} - p_i^*}{p_i^*} - \left( \frac{\hat{p}_{-i} - p_i^*}{p_i^*} \right)^2 + ... \right) \\
&= -\frac{2}{n\left(n-1\right)} \sum_i \sum_{l \neq i} u_i u_l L_{il} \frac{1}{\left(p_i^*\right)^2} \\
&\quad \cdot \left\{ \left\{ \sum_{s=1}^{r_1} \bar{\lambda}_s \sum_{\bar{x}_1 \in D, \left\| \bar{X}_i - \bar{x}_1 \right\|_s = 1} \bar{p}\left(\bar{x}_1\right) + \sum_{s,t} \bar{\lambda}_s \bar{\lambda}_t \sum_{\bar{x}_1 \in D, \left\| \bar{X}_i - \bar{x}_1 \right\|_{s,t} = 1} \bar{p}\left(\bar{x}_1\right) \right\} \right. \\
&\quad \left. \cdot E_1 \left[ \tilde{L} \left( \tilde{X}_1, \tilde{X}_i \right) \right] + O\left( n^{-1/2} \right) + o\left( \bar{\lambda}^2 \right) \right\} \\
&= -\sum_{s=1}^{r_1} \bar{\lambda}_s \left\{ \frac{2}{n\left(n-1\right)} \sum_i \sum_{l>i} u_i u_l I_{\left\| \bar{X}_l - \bar{X}_i \right\| = 0} \tilde{L}_{il} \frac{1}{p_i^*} \right. \\
&\quad \left. \cdot \frac{1}{\bar{p}_i} \sum_{\bar{x}_1 \in D, \left\| \bar{X}_i - \bar{x}_1 \right\|_s = 1} \bar{p}\left(\bar{x}_1\right) \right\} + O\left( n^{-1} \bar{\lambda}^2 \right) + O\left( n^{-3/2} \right) \qquad (2.78)
\end{aligned}
$$

where the second, the third equality holds since (2.33), (2.32) respectively.

So, by (2.77), (2.78)

$$\begin{aligned}
T_2 &= \frac{4}{n(n-1)}\sum_i\sum_{l>i} u_i u_l I_{\|\bar{X}_l-\bar{X}_i\|=0}\tilde{L}_{il}/\left(\bar{p}_i E_1\left[\tilde{L}\left(\tilde{X}_1,\tilde{X}_i\right)\right]\right) \\
&\quad +\sum_{s=1}^{r_1}\bar{\lambda}_s\left\{\frac{4}{n(n-1)}\sum_i\sum_{l>i} u_i u_l\cdot I_{\left\|\bar{X}_l-\bar{X}_i\right\|_s=1}\cdot\tilde{L}_{il}/\left(\bar{p}_i E_1\left[\tilde{L}\ \left(\tilde{X}_1,\tilde{X}_i\right)\right]\right)\right\} \\
&\quad -\sum_{s=1}^{r_1}\bar{\lambda}_s\frac{2}{n(n-1)}\sum_i\sum_{l>i} u_i u_l I_{\|\bar{X}_l-\bar{X}_i\|=0}\tilde{L}_{il}\frac{1}{p_i^*} \\
&\quad \cdot\frac{1}{\bar{p}_i}\sum_{\bar{x}_1\in D,\left\|\bar{X}_i-\bar{x}_1\right\|_s=1}\bar{p}(\bar{x}_1)+O\left(\frac{1}{n}\bar{\lambda}^2\right)+O\left(n^{-3/2}\right) \\
&= \frac{4}{n(n-1)}\sum_i\sum_{l>i} u_i u_l I_{\|\bar{X}_l-\bar{X}_i\|=0}\tilde{L}_{il}/\left(\bar{p}_i E_1\left[\tilde{L}\left(\tilde{X}_1,\tilde{X}_i\right)\right]\right) \\
&\quad +O\left(\frac{1}{n}\bar{\lambda}\right)+O\left(n^{-3/2}\right) \qquad (2.79)
\end{aligned}$$

$$\begin{aligned}
T_1 &= \frac{2}{n(n-1)}\sum_i\sum_{l\neq i} u_i(g_i-g_l)L_{il}/\hat{p}_{-i} \\
&= \frac{2}{n(n-1)}\sum_i\sum_{l\neq i} u_i(g_i-g_l)L_{il}/p_i^* \\
&\quad +\frac{2}{n(n-1)}\sum_i\sum_{l\ \neq i} u_i(g_i-g_l)L_{il}\left(\frac{1}{\hat{p}_{-i}}\ -\frac{1}{p_i^*}\right) \\
&= \frac{2}{n(n-1)}\sum_i\sum_{l\neq i} u_i(g_i-g_l)L_{il}/p_i^* \\
&\quad -\frac{2}{n(n-1)}\sum_i\sum_{l\ \neq i} u_i(g_i-g_l)L_{il}\frac{1}{p_i^*}\ \left(\frac{\hat{p}_{-i}-p_i^*}{p_i\ ^*}-\left(\frac{\hat{p}_{-i}-p_i^*}{p_i\ ^*}\right)^2+...\right) \\
&= \Upsilon_1-\Upsilon_2 \qquad (2.80)
\end{aligned}$$

$$\begin{aligned}
\Upsilon_1 &= \frac{2}{n(n-1)}\sum_i\sum_{l\neq i} u_i(g_i-g_l)\bar{L}_{il}/\bar{p}_i\cdot\frac{\tilde{L}_{il}}{E_1\left[\tilde{L}\left(\tilde{X}_1,\tilde{X}_i\right)\right]} \\
&= \frac{2}{n(n-1)}\sum_i\sum_{l\neq i} u_i(g_i-g_l)/\bar{p}_i\cdot\frac{\tilde{L}_{il}}{E_1\left[\tilde{L}\left(\tilde{X}_1,\tilde{X}_i\right)\right]}[\sum_{s=1}^{r_1}\bar{\lambda}_s I_{\left\|\bar{X}_l-\bar{X}_i\right\|_s=1} \\
&\quad +\sum_{s,t}\bar{\lambda}_s\bar{\lambda}_t I_{\left\|\bar{X}_l-\bar{X}_i\right\|_{s,t}=1}+...] \\
&= \sum_{s=1}^{r_1}\bar{\lambda}_s\frac{2}{n(n-1)}\sum_i\sum_{l\neq i} u_i(g_i-g_l)/\bar{p}_i\cdot\frac{\tilde{L}_{il}}{E_1\left[\tilde{L}\left(\tilde{X}_1,\tilde{X}_i\right)\right]}I_{\left\|\bar{X}_l-\bar{X}_i\right\|_s=1} \\
&\quad +O_p\left[\sqrt{\frac{1}{n}}\bar{\lambda}^2\right] \qquad (2.81)
\end{aligned}$$

Obviously

$$\Upsilon_2 = O_p\left(\frac{1}{n}\bar{\lambda}\right)+O_p\left(\sqrt{\frac{1}{n}}\bar{\lambda}^2\right) \qquad (2.82)$$

Hence, by (2.80), (2.81) and (2.82) we have

$$\begin{aligned}
T_1 &= \sum_{s=1}^{r_1}\bar{\lambda}_s\left\{\frac{2}{n(n-1)}\sum_i\sum_{l\neq i} u_i(g_i-g_l)/\bar{p}_i I_{\left\|\bar{X}_l-\bar{X}_i\right\|_s=1}\cdot\frac{\tilde{L}_{il}}{E_1\left[\tilde{L}\left(\tilde{X}_1,\tilde{X}_i\right)\right]}\right\} \\
&\quad +O_p\left[\sqrt{\frac{1}{n}}\bar{\lambda}^2\right]+O_p\left(\frac{1}{n}\bar{\lambda}\right) \qquad (2.83)
\end{aligned}$$

By (2.76), (2.79) and (2.83), we obtain

$$\begin{aligned}
I_2 &= \sum_{s=1}^{r_1}\bar{\lambda}_s\left\{\frac{2}{n(n-1)}\sum_i\sum_{l\neq i} u_i(g_i-g_l)/\bar{p}_i I_{\left\|\bar{X}_l-\bar{X}_i\right\|_s=1}\cdot\frac{\tilde{L}_{il}}{E_1\left[\tilde{L}\left(\tilde{X}_1,\tilde{X}_i\right)\right]}\right\} \\
&\quad -\frac{4}{n(n-1)}\sum_i\sum_{l>i} u_i u_l I_{\left\|\bar{X}_l-\bar{X}_i\right\|=0}\tilde{L}_{il}/\left(\bar{p}_i E_1\left[\tilde{L}\left(\tilde{X}_1,\tilde{X}_i\right)\right]\right) \\
&\quad +O_p\left[\sqrt{\frac{1}{n}}\bar{\lambda}^2\right]+O_p\left(\frac{1}{n}\bar{\lambda}\right)+O_p\left(n^{-3/2}\right) \qquad (2.84)
\end{aligned}$$

# CHAPTER III

# CROSS-VALIDATION AND THE ESTIMATION OF PROBABILITY DISTRIBUTIONS WITH CATEGORICAL DATA

## A. Introduction

It is well known that the traditional 'frequency-based' nonparametric approach may be used to consistently estimate a joint probability distribution defined over categorical variables. In this chapter we focus our attention on an alternative nonparametric approach when confronted with discrete variables where, instead of using the conventional 'frequency' method to handle the discrete variables, we choose to *smooth* the discrete variables. There is a growing literature on nonparametric smoothing of discrete variables dating back to the pioneering work of Aitchison and Aitken (1976); see Aerts, Augustyns and Janssen (1997a, 1997b), Ahmad and Cerrito (1994), Bowman (1980), Grund (1993), Hall (1981), Hall, Racine and Li (2004), Izenman (1991), and Li and Racine (2003), Simonoff (1983), and Tutz (1991), to name but a few (see also the monographs by Hart (1997), Scott (1992) and Simonoff (1996)).

From a statistical perspective, the kernel smoothing of discrete variables may introduce some finite sample bias; however, it also reduce the variance, leading to a reduction in the mean square error of the resulting estimator relative to the frequency-based estimator. More importantly, we show that the kernel smoothing method, particularly when used in conjunction with a data-driven method of bandwidth selection such as cross-validation, has the ability to smooth over the 'uniformly distributed' variables (a specific definition of 'uniformly distributed variables' is given in Section 2). When uniformly distributed variables are encountered, the conventional frequency method still splits the sample into many discrete cells, including those cells arising

from the presence of the uniformly distributed variables. In contrast, the smoothing-cross-validation method can automatically oversmooth these variables with a high probability, thereby reducing the dimension of the nonparametric model to that associated with the non-uniformly distributed variables only. In such cases, we reduce estimation variance substantially without increasing bias, and we often obtain impressive efficiency gains.

As we will show in Section C, the asymptotic behavior of the smoothing parameters depends on whether a variable is uniformly distributed or not. We emphasize here that it is important to explicitly consider the vector-valued smoothing parameter case. If one were to use a scalar smoothing parameter (i.e.,the same value for each variable), as is often done for the convenience of the theoretical analyses (e.g., Li and Racine (2003)), then it would not be possible to benefit from the differing asymptotic behavior alluded to above. Thus, our results extend Racine and Li (2004) to the practically relevant vector-valued smoothing parameter case, allowing for the existence of uniformly distributed variables, which results in differing asymptotic behavior of the smoothing parameters.

## B. Smooth Estimation of Joint Distributions with Discrete Data

In this section we consider the estimation of a probability function defined over discrete data $X$, where $X$ is an $r$-dimensional discrete random vector taking values on $\mathcal{S}$, the support of $X$. We use $x^s$ and $X_i^s$ to denote the $s$th component of $x$ and $X_i$ $(i = 1, \ldots, n)$, respectively. Following Aitchison and Aitken (1976), for $x^s$, $X_i^s \in \mathcal{S}_s = \{0, 1, \ldots, c_s - 1\}$ ($x^s$ takes $c_s$ different values), we define a univariate kernel function

$$l(X_i^s, x^s, \lambda_s) = \begin{cases} 1 - \lambda_s & \text{if } X_i^s = x^s, \\ \lambda_s/(c_s - 1) & \text{if } X_i^s \neq x^s. \end{cases} \tag{3.1}$$

The range of $\lambda_s$ is $[0,(c_s-1)/c_s]$. Note that when $\lambda_s=0$, $l(X_i^s,x^s,0)=I(X_i^s=x^s)$ becomes an indicator function. Here we use $I(.)$ to denote an indicator function, that is, $I(A)=1$ if event $A$ is true, zero otherwise. If $\lambda_s=(c_s-1)/c_s$, then $l(X_i^s,x^s,\frac{c_s-1}{c_s})=1/c_s$ is a constant for *all* values of $X_i^s$ and $x^s$.

We shall follow the product kernel convention, where the product kernel function is given by

$$L(X_i,x,\lambda)=\prod_{t=1}^{r}l(X_i^s,x^s,\lambda_s)=\prod_{s=1}^{r}\{\lambda_s/(c_s-1)\}^{I_{x_i^s\neq x^s}}(1-\lambda_s)^{I_{x_i^s=x^s}}, \tag{3.2}$$

where $I_{x_i^s\neq x^s}=I(X_i^s\neq x^s)$, and $I_{x_i^s=x^s}=I(X_i^s=x^s)$.

For a given value of the smoothing parameter vector $\lambda=(\lambda_1,\ldots,\lambda_r)$, we estimate $p(x)$ by

$$\hat{p}(x)=\frac{1}{n}\sum_{i=1}^{n}L(X_i,x,\lambda). \tag{3.3}$$

Note that the weight function $l(.,.,.)$ defined in (3.1) adds up to one, which ensures that $\hat{p}(x)$ defined in (3.3) is a proper probability measure, i.e., $\sum_{x\in\mathcal{S}}\hat{p}(x)=1$. Note also that if $\lambda_s=0$ for all $1\leq s\leq r$, then (3.3) collapses to the conventional frequency estimator. Therefore, the smoothed estimator $\hat{p}(x)$ nests the frequency estimator as a special case. Another interesting case arises when $\lambda_s$ assumes its upper bound value $(c_s-1)/c_s$, since then $\hat{p}(x)$ becomes unrelated to $x^s$. In this case we say that $x^s$ is completely 'smoothed out', and the resulting estimator $\hat{p}(x)$ becomes unrelated to $x^s$ in the sense that, when $\lambda_s=(c_s-1)/c_s$,

$$\hat{p}(x^1,\ldots,x^{s-1},x^s,x^{s+1},\ldots,x^r)=\hat{p}(x^1,\ldots,x^{s-1},z^s,x^{s+1},\ldots,x^r), \tag{3.4}$$

for all $x^s,z^s\in\{0,1,\ldots,c_s-1\}$, and $x^{-s}\in\mathcal{S}_{-s}$, where $x^{-s}=(x^1,\ldots,x^{s-1},x^{s+1},\ldots,x^r)$ denotes $x$ with $x^s$ removed, and where $\mathcal{S}_{-s}$ is the support of $X^{-s}$. With $\lambda_s$ assuming its upper extreme value $(c_s-1)/c_s$, we obtain the best possible choice for $\lambda_s$ when $x^s$

is indeed *uniformly distributed* as defined by

$$p(x^1, \ldots, x^{s-1}, x^s, x^{s+1}, \ldots, x^r) = p(x^1, \ldots, x^{s-1}, z^s, x^{s+1}, \ldots, x^r), \tag{3.5}$$

for all $x^s, z^s \in \{0, 1, \ldots, c_s - 1\}$ and $x^{-s} \in \mathcal{S}_{-s}$.

A comparison of (3.4) with (3.5) reveals that, when $\lambda_s = (c_s - 1)/c_s$, the estimator $\hat{p}(x)$ satisfies the uniform distribution property (for $x^s$). This is a highly desirable property since, when $x^s$ is a uniformly distributed variable, $p(x)$ is invariant to $x^s$ and one should smooth out $x^s$ rather than splitting the sample into different discrete cells that would otherwise be generated by $x^s$ when using the conventional frequency estimator. In Section 5 we show via simulations that our cross-validation method can indeed select $\lambda_s = (c_s - 1)/c_s$ with a high probability when $x^s$ is uniformly distributed, and in this case $\hat{p}(x)$ is much more efficient than the frequency estimator, which uses $\lambda_s = 0$.

As was the case when dealing with continuous variables, the selection of smoothing parameters is of crucial importance. We suggest choosing the smoothing parameters $\lambda_1, \ldots, \lambda_r$ by minimizing the squared difference between $\hat{p}(\cdot)$ and $p(\cdot)$, which is given by ($\sum_x \equiv \sum_{x \in \mathcal{S}}$)

$$\begin{aligned} J_n &= \sum_x [\hat{p}(x) - p(x)]^2 \\ &= \sum_x [\hat{p}(x)]^2 - 2\sum_x \hat{p}(x)p(x) + \sum_x [p(x)]^2, \\ &\equiv J_{1n} - 2J_{2n} + \sum_x [p(x)]^2, \end{aligned} \tag{3.6}$$

where $J_{1n} = \sum_x [\hat{p}(x)]^2$ and $J_{2n} = \sum_x \hat{p}(x)p(x)$. Note that $J_{2n} = E[\hat{p}(X)]$, hence we can estimate $J_{2n}$ by replacing the unknown population mean with the sample mean

to obtain

$$\hat{J}_{2n} = \frac{1}{n}\sum_{i=1}^{n}\hat{p}_{-i}(X_i) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} L_{ij}, \tag{3.7}$$

where $L_{ij} = L(X_i, X_j, \lambda)$ and $\hat{p}_{-i}(X_i) = (n-1)^{-1}\sum_{j=1,j\neq i}^{n} L_{ij}$ is the leave-one-out kernel estimator of $p(X_i)$. The last term on the right-hand-side of (3.6) is unrelated to $\lambda_1, \ldots, \lambda_r$. Therefore, we choose $\lambda_1, \ldots, \lambda_r$ to minimize

$$CV(\lambda) = J_{1n} - 2\hat{J}_{2n} = \frac{1}{n}\sum_{x}[\hat{p}(x)]^2 - \frac{2}{n(n-1)}\sum_{i}\sum_{j\neq i} L_{ij}. \tag{3.8}$$

Let $\hat{\lambda}_1,\ldots,\hat{\lambda}_r$ denote the cross-validation selection of $\lambda_s$'s that minimize (3.8). In subsection F.1. we show that, when $x^s$ does not have a uniform marginal distribution, then $\hat{\lambda}_s = o_p(1)$. Note this result is quite different compared with the cross-validation estimation of a density function with continuous variables, where one has to assume that smoothing parameter, say $h$, to take values in a shrinking set (i.e., $h \to 0$ as $n \to \infty$) in order to derive the rate of convergence of the cross-validation selected smoothing parameter, say $\hat{h}$ (e.g., $\hat{h} \sim n^{-1/5}$ in the univariate case). Here we do not impose the restriction that $\hat{\lambda}_s$ taking values in a shrinking set, instead we prove that $\hat{\lambda}_s = o_p(1)$ provided that $x^s$ does not have a uniform marginal distribution. Let $\lambda_{(r)} = (\lambda_1, \ldots, \lambda_r)'$ denote the $r \times 1$ vector of smoothing parameters. We further show that

$$CV(\lambda) = \lambda_{(r)}'\Omega\lambda_{(r)} - n^{-1}c_n'\lambda_{(r)} + o_p\left(n^{-1}\sum_{s=1}^{r}\lambda_s + \sum_{s=1}^{r}\lambda_s^2\right) + \text{ terms not related to } \lambda, \tag{3.9}$$

where $\Omega$ is an $r \times r$ positive definite matrix, and $c_n$ is a $O_p(1)$ random vector of dimension $r \times 1$. The explicit expressions for $\Omega$ and $c_n$ are given in subsection F.1. Since $\Omega$ is positive definite, minimizing (3.9) over $\lambda_{(r)}$ leads to

$$\hat{\lambda}_{(r)} = n^{-1}\Omega^{-1}c_n + o_p(n^{-1}). \tag{3.10}$$

(3.10) shows that $\hat{\lambda}_s = O_p(n^{-1})$ for $s = 1, \ldots, r$. The crucial condition needed for (3.10) to hold is that $\Omega$ is positive definite. We show in subsection F.1. that if $x^s$ does not have a uniform marginal distribution for all $s = 1, \ldots, r$, then $\Omega$ is positive definite. We summarize the above result in the following theorem.

**Theorem III.1** *Assuming that $x^s$ does not have a uniform marginal distribution for all $s = 1, \ldots, r$, then*

$$\hat{\lambda}_s = O_p(n^{-1}) \text{ for } s = 1, \ldots, r.$$

Proof: In subsection F.1. we first show that $\hat{\lambda}_s = o_p(1)$ for all $s = 1, \ldots, r$ so that we can expand $CV(\lambda)$ so as to yield leading terms that are low order polynomials in $\lambda_s$, $s = 1 \ldots, q$ (i.e., (3.9)). We then show that $\Omega$ is positive definite. Hence, Theorem III.1 follows from (3.9).

If $\lambda_s = o(1)$ and is non-stochastic ($s = 1, \ldots, r$), then it is straightforward to show that

$$\begin{aligned} E[\hat{p}(x)] &= p(x) + \sum_{s=1}^{r} [p_{s,1}(x) - p(x)]\lambda_s + O\left(\sum_{s=1}^{r} \lambda_s^2\right), \\ var(\hat{p}(x)) &= n^{-1} p(x)(1 - p(x)) + O\left(n^{-1} \sum_{s=1}^{r} \lambda_s\right), \end{aligned} \tag{3.11}$$

where the $p_{s,1}(x) = \frac{1}{c_s - 1} \sum_{z^s \neq x^s, z^{-s} = x^{-s}} p(z)$. With the fast $n^{-1}$ rate of convergence of the $\hat{\lambda}_s$'s given in Theorem III.1, the following result follows from (3.11).

**Theorem III.2** *Let $\hat{p}(x)$ be defined as in (3.3) with $\lambda$ replaced by $\hat{\lambda}$. Then under the same conditions as in Theorem III.1, we have*

*(i)* $\hat{p}(x) - p(x) = O_p(n^{-1/2})$,

*(ii)* $n^{1/2}(\hat{p}(x) - p(x)) \xrightarrow{d} N(0, p(x)(1 - p(x)))$.

Proof: Theorem III.2 follows from Theorem III.1 and (3.11).

Theorem III.2 reveals that the *asymptotic* distribution of $\hat{p}(x)$ is the same as that of the frequency estimator. This is exactly what one should expect since, asymptotically, we have an infinite sample size; hence even with sample splitting we still have an infinite amount of data lying in each cell. Therefore, if the sample size is sufficiently large, $\hat{\lambda}_1, \ldots, \hat{\lambda}_r$ should all be close to zero, and our cross-validation estimator $\hat{p}(x)$ will be quite close to the frequency estimator. However, in finite sample applications the two approaches can yield very different results.

In the next section we will show that when $x^s$ is uniformly distributed in the sense of (3.5), then $\hat{\lambda}_s$ will not converge to zero; rather it will assume its upper extreme value $(c_s - 1)/c_s$ with a high probability. Careful readers may notice that there is still a case not covered here. That is, $x^s$ may have a marginal uniform distribution and at the same time be non-uniformly distributed (i.e., not satisfy (3.5)). For example, Table VIII shows one such example whereby, while $x^1$ and $x^2$ are both non-uniformly distributed, they both have uniform marginals: $p_j(x^j) = 1/2$ for $j = 1, 2$ ($a \in (0, 1/2)$).

Table VIII. A Non-uniformly Distributed Random Variable with Marginal Distributions

| $P(x^1, x^2)$ | $x^1 = 0$ | $x^1 = 1$ |
|---|---|---|
| $x^2 = 0$ | $a$ | $1/2 - a$ |
| $x^2 = 1$ | $1/2 - a$ | $a$ |

Even though Theorem III.1 does not cover this case, intuitively one would expect that $\hat{\lambda}_s \to 0$ $(s = 1, 2)$ as $n \to \infty$, for otherwise it would lead to inconsistent estimation results. This is indeed true as we show in the next theorem. For notational simplicity we will restrict attention to the two-variable case.

**Theorem III.3** *Assume that both $x^1$ and $x^2$ are non-uniformly distributed variables both having uniform marginal distributions, while all other $x^s$ have non-uniform marginals ($s = 3, \ldots, r$). Then*

$\hat{\lambda}_s = O_p(n^{-1})$ *for all* $s = 1, \ldots, r$.

The proof of Theorem III.3 is given in subsection F.1.

C. The Uniformly Distributed Variable Case

In Section 2 we considered the case where none of the components of $X$ was uniformly distributed. In this section we allow for the presence of a uniformly distributed variable. Without loss of generality, we assume that the first $r-1$ components of $x$, denoted $\bar{x} = (x^1, \ldots, x^{r-1})$, are non-uniformly distributed, while the last one, the $r$th component of $x$ denoted $\tilde{x} = x^r$, is uniformly distributed in the sense of (3.5). Let $\bar{p}(\bar{x})$ denote the marginal probability function of $\bar{x}$. Summing over $\tilde{x}$ and using (3.5) we get

$$p(\bar{x}, \tilde{x}) = \bar{p}(\bar{x})/c_r \text{ for all } \tilde{x} \in \tilde{S}, \tag{3.12}$$

where $\tilde{\mathcal{S}} = \{0, 1, \ldots, c_r - 1\}$ is the support of $\tilde{X}$. (3.12) shows that $p(x)$ varies only with $\bar{x}$, but not with $\tilde{x}$. When $X$ is uniformly distributed with respect to $\tilde{x}$ and one chooses $\lambda_r = (c_r - 1)/c_r$, then the resulting $\hat{p}(x)$ effectively uses this information, does not introduce bias (related to $\tilde{x}$), and reduces the variance significantly.

The above arguments underscore the desirability of a data-driven bandwidth selection method having the ability to select a large value of $\lambda_s$ if $x^s$ is a uniformly distributed variable, and a small value for $\lambda_s$ otherwise. It turns out that least squares cross-validation indeed has this desired property. The following theorem covers this case.

**Theorem III.4** *Assuming that $x^1, \ldots, x^{r-1}$ are non-uniformly distributed variables and $x^r$ is uniformly distributed, then*

*(i) $\hat{\lambda}_s = O_p(n^{-1}) = o_p(1)$ for $s = 1, \ldots, r-1$, and*

*(ii) asymptotically there is a positive constant $\delta \in (0,1)$ for which $CV(\lambda)$ is minimized at $\hat{\lambda}_r = (c_r - 1)/c_r$ with probability $\delta$.*

Theorem III.4 is proved in subsection F.2. Theorem III.4 states that the smoothing parameter associated with the uniformly distributed variable will not converge to zero, but rather tends to assume its upper extreme value with a high probability, so that the estimated probability function satisfies the uniform distribution condition of (3.5) (for $x^r$). In this case, $\hat{p}(x)$ is more efficient than the frequency estimator, which does not impose this restriction. Given the proofs in subsection F.1., the difficulty of determining the exact value of $\delta$ can be appreciated. For a range of the data generating processes considered in Section 5, simulations reveal that $\delta$ appears to lie between 0.60 to 0.64.

The conclusion of Theorem III.4 is qualitatively different from the conditional density estimation results for mixed categorical and continuous variables. Hall, Racine, and Li (2003) consider the cross-validated kernel estimation of conditional density functions with mixed discrete and continuous variables. In that context, irrelevant variables are those conditional variables that are orthogonal to the dependent variable. Hall et al. (2003) show that the cross-validation method has the ability to automatically remove the irrelevant variables in the sense that $\lambda_s \to (c_s - 1)/c_s$ in probability when $x^s$ is an irrelevant discrete conditioning variable. Here Theorem III.4 claims that, while there is a positive probability $\delta$ of $\lambda_s$ assuming its upper extreme value of $(c_s - 1)/c_s$, there is also a positive probability $1 - \delta$ that $\lambda_s$ takes values in $[0, (c_s - 1)/c_s)$. Thus, the discrete-variable-only case with uniformly distributed

variables is markedly different from the case of conditional density estimation with mixed variables when there exist irrelevant variables.

When there exists more than one uniformly distributed random variable, similar results hold. That is, $\hat{\lambda}_s = O_p(n^{-1})$ if $x^s$ is non-uniformly distributed, and $\hat{\lambda}_s$ has a positive probability of assuming its upper extreme value $(c_s - 1)/c_s$ if $x^s$ is a uniformly distributed variable. Since the analysis is quite tedious for the general case, we will rely instead on simulations to investigate this case in the next section.

## D. Monte Carlo Simulations

In this section we report some simulations designed to examine the finite sample performance of the proposed CV method. We consider the following cases:

Case (i): we first consider a simple case having one non-uniform variable and one uniform variable, both of which are binary: $p(x^1, x^2) = p_1(x^1)/2$ with $p_1(x^1 = 0) = 0.3$ and $p_1(x^1 = 1) = 0.7$ ($x^2$ is a uniformly distributed binary variable).

Case (ii): the same as in case (i) above except that now $x^2$ can assume four different values so that $p(x^1, x^2) = p_1(x^1)/4$ ($x^2$ is a uniformly distributed variable taking values in $\{0, 1, 2, 3\}$).

Case (iii): similar to case (i) but with an additional uniformly distributed variable. $p(x^1, x^2, x^3) = p_1(x^1)/4$ where $p_1(x^1)$ is the same as in case (i) ($x^2$ and $x^3$ are both uniformly distributed binary variables).

We examine the cross-validated values of $\lambda$. Our theory predicts that $\hat{\lambda}_1 \to 0$ in probability, while $\hat{\lambda}_2$ (and $\hat{\lambda}_3$ for case (iii)) has a tendency to take its upper bound value of 1/2 (3/4 for case (ii)). We compute the Average MSE by $\sum_{j=1}^{2000} \sum_x [\hat{p}_{(j)}(x) - p(x)]^2$, where $\hat{p}_{(j)}(x)$ is the estimated $p(x)$ in the $j$th simulation. We also compute a conventional frequency estimator that uses $\lambda_1 = \lambda_2 = 0$ ($\lambda_3 = 0$) for estimating $p(x)$.

The number of simulations is 1,000, and the sample sizes are $n = 50$, 100 and 200.

The MSE results for case (i) are reported in Table IX, and we define Rel. MSE $= MSE_{cv}/MSE_{freq}$. We observe that our cross-validation based method yields a much smaller finite sample MSE than the conventional frequency estimator when there exists a uniformly distributed variable. As we explained earlier, our CV-based estimator has a tendency to select $\hat{\lambda}_2 = 1/2$, its upper extreme value. When $\hat{\lambda}_2 = 1/2$, our estimator effectively uses the information that $p(x) = p_1(x^1)/2$ and we do not split the sample for different values of $x^2$. Thus, our estimator is expected to have better finite sample performance than the frequency estimator.

Table IX. MSE for Case (i)

| $n$ | $MSE_{cv}$ | $MSE_{freq}$ | Rel. MSE |
|---|---|---|---|
| 50 | 0.001177 | 0.002900 | 0.405945 |
| 100 | 0.000465 | 0.001400 | 0.332301 |
| 200 | 0.000245 | 0.000727 | 0.336838 |

Table X. MSE for Case (ii)

| $n$ | $MSE_{cv}$ | $MSE_{freq}$ | Rel. MSE |
|---|---|---|---|
| 50 | 0.000294 | 0.002171 | 0.135571 |
| 100 | 0.000175 | 0.001041 | 0.168091 |
| 200 | 0.000061 | 0.000499 | 0.122766 |

Table X reports the estimated MSE for case (ii). We observe that, compared with the results for case (i), the CV method produces an even better relative MSE than the frequency method. This is because the uniformly distributed variable $x^2$ now assumes four different values, and since we do not split the sample into different discrete cells generated by $x^2$, the relative performance of our CV method improves as the uniformly distributed variable assumes an increasing number of values. For $n = 50$, 100, and 200, the percentages of $\hat{\lambda}_2$ that exceed 0.749 (i.e., taking values in $(0.7499, .0.75])$,[1] ares 0.601, 0.600, and 0.606, respectively.

Table XI. MSE for Case (iii)

| $n$ | $MSE_{cv}$ | $MSE_{freq}$ | Rel. MSE |
|---|---|---|---|
| 50 | 0.000495 | 0.002199 | 0.225162 |
| 100 | 0.000197 | 0.001064 | 0.185539 |
| 200 | 0.000086 | 0.000518 | 0.166466 |

Finally, Table XI provides the estimated MSE for case (iii). We observe that when there exists more than one uniformly distributed variable, the relative efficiency gains of our CV estimator over the frequency estimator further improves compared with case (i). For case (iii) $x^2$ and $x^3$ are uniformly distributed, and both $\hat{\lambda}_2$ and $\hat{\lambda}_3$ have roughly a 60% probability of assuming their upper extreme value, 1/2.

[1]Note that since $c_2 = 4$, the upper bound value for $\lambda_2$ is $(4 - 1)/4 = 0.75$.

## E. Conclusion

We analyze the problem of estimating a joint distribution defined over a set of discrete variables. We use a smoothing kernel estimator to estimate the joint distribution, allow for the case in which some of the discrete variables are uniformly distributed, and explicitly address the vector-valued smoothing parameter case. The cross-validated smoothing parameters differ in their asymptotic behavior depending on whether a variable is uniformly distributed or not. Simulation experiments highlight how the proposed estimator performs much better than the commonly used frequency estimator.

## F. Proofs

### 1. Proof of Theorem III.3

We first prove that $\hat{\lambda}_s = o_p(1)$ for all $s = 1, \ldots, r$.

**Lemma III.1** *$\hat{\lambda}_s = o_p(1)$ for all $s = 1, \ldots, r$.*

Proof: Recall that $J_n(\lambda) = \sum_x [\hat{p}(x)]^2 - 2E[\hat{p}(x)] + \sum_x p(x)^2$ and define $\hat{J}_n(\lambda) = \sum_x [\hat{p}(x)]^2 - 2n^{-1}\sum_{i=1}^n \hat{p}_{-i}(X_i) + \sum_x p(x)^2$. Obviously, $n^{-1}\sum_{i=1}^n \hat{p}_{-i}(X_i) = E[\hat{p}_{-i}(X_i)] + o_p(1)$ uniformly in $\lambda$. So $\hat{J}_n(\lambda) = J_n(\lambda) + o_p(1)$.

Next, $0 \leq J_n(\hat{\lambda}) = \hat{J}_n(\hat{\lambda}) + o_p(1)$, and $\hat{J}_n(\hat{\lambda}) \leq \hat{J}_n(0) = o_p(1)$ because $\lambda = 0$ corresponds to the frequency estimator and it is well established that $\hat{J}_n(0) = O_p(n^{-1/2}) = o_p(1)$. Thus, $\hat{J}_n(\hat{\lambda}) = o_p(1)$. And since $\hat{J}_n(\lambda) = J_n(\lambda) + o_p(1)$, we get $J_n(\hat{\lambda}) = o_p(1)$.

From $J_n(\hat{\lambda}) = \sum_x [\hat{p}(x) - p(x)]^2 = o_p(1)$, we know that $\hat{p}(x) - p(x) = o_p(1)$ for all $x \in \mathcal{S}$, i.e.,

$$n^{-1}\sum_i L(x_i, x, \hat{\lambda}) - p(x) = o_p(1) \text{ for all } x \in \mathcal{S}. \tag{3.13}$$

Summing (3.13) over $x^2, \ldots, x^r$ we obtain

$$n^{-1} \sum_i l(x_i^1, x^1, \lambda_1) - p_1(x^1) = o_p(1) \text{ for all } x^1 \in \mathcal{S}_1, \tag{3.14}$$

where $\mathcal{S}_1 = \{0, 1, \ldots, c_1 - 1\}$ is the support of $X^1$.

Note that $l(x_i^1, x^1, \hat{\lambda}_1) = (1 - \hat{\lambda}_1) I_{x_i^1 = x^1} + \frac{\hat{\lambda}_1}{c_1 - 1} I_{x_i^1 \neq x^1}$. So (3.14) is

$$\begin{aligned}
o_p(1) &= n^{-1} \sum_i I_{x_i^1 = x^1} + \hat{\lambda}_1 n^{-1} \sum_i \left[ \frac{1}{c_1 - 1} I_{x_i^1 \neq x^1} - I_{x_i^1 = x^1} \right] - p_1(x^1) \\
&= \hat{\lambda}_1 n^{-1} \sum_i \left[ \frac{1}{c_1 - 1} I_{x_i^1 \neq x^1} - I_{x_i^1 = x^1} \right] + O_p(n^{-1/2}) \\
&= \hat{\lambda}_1 \left[ \frac{1}{c_1 - 1} \sum_{z^1 \neq x^1} p_1(z^1) - p_1(x^1) \right] + O_p(n^{-1/2}) \\
&= \hat{\lambda}_1 \frac{c_1}{c_1 - 1} \left[ \frac{1}{c_1} - p_1(x^1) \right] + O_p(n^{-1/2}),
\end{aligned} \tag{3.15}$$

for all $x^1 \in \mathcal{S}_1$. Since $p_1(x^1) \neq 1/c_1$ for at least some $x^1 \in \mathcal{S}_1$, we know from (3.15) that $\hat{\lambda}_1 = o_p(1)$. By symmetry we have $\hat{\lambda}_s = o_p(1)$ for all $s = 1, \ldots, r$.

This completes the proof for Lemma III.1.

Given that $\hat{\lambda}_s = o_p(1)$, we next expand $CV(\lambda)$ in polynomials of $\lambda_s$ and show that $\hat{\lambda}_s = O_p(n^{-1})$ for $s = 1, \ldots, r$. Defining $L_{ij}^{(2)} = \sum_{x \in \mathcal{S}} L_{x,x_i} L_{x,x_j}$, then $CV(\lambda)$ can be written as

$$\begin{aligned}
CV(\lambda) &= n^{-2} \sum_{i=1}^n \sum_{j=1}^n L_{ij}^{(2)} - \frac{2}{n(n-1)} \sum_i \sum_{j \neq i} L_{ij} \\
&= n^{-2} \sum_{i=1}^n L_{ii}^{(2)} + \frac{1}{n(n-1)} \sum_i \sum_{j \neq i} (L_{ij}^{(2)} - 2L_{ij}) - \frac{1}{n^2(n-1)} \sum_i \sum_{j \neq i} L_{ij}^{(2)} \\
&\equiv I_{1n} + I_{2n} - I_{3n},
\end{aligned} \tag{3.16}$$

where the definition of $I_{jn}$ $(j = 1, 2, 3)$ should be apparent.

We will use $O(\Lambda^l)$ to denote terms either of order $O(\sum_{s=1}^r \lambda^l)$ or terms unrelated to $\lambda$. We define the following notation for some indicator functions that follow:

$$\begin{aligned} I_{1,x,z} &= I(\text{ if } x \text{ and } z \text{ differ by exactly one component}), \\ I_{2,x,z} &= I(\text{ if } x \text{ and } z \text{ differ by exactly two components}), \\ I^s_{1,x,z} &= I(x^s \neq z^s) \prod_{t \neq s} I(x^t = z^t), \\ I^{u,v}_{2,x,z} &= I(x^u \neq z^u) I(x^v \neq z^v) \prod_{s \neq, u, v} I(x^s = z^s). \end{aligned} \tag{3.17}$$

Note that $I_{l,x,z} = 1$ if and only if $x$ and $z$ differ by exactly $l$ components ($l = 1, 2$), $I^s_{1,x,z} = 1$ if and only if $x$ and $z$ differ only in the $s$th component, and $I^{u,v}_{2,x,z} = 1$ if and only if $x$ and $z$ differ by two components, the $u$th and the $v$th components.

Also, define

$$\begin{aligned} p_{s,1}(x) &= \frac{1}{c_s - 1} \sum_z p(z) I^s_{1,z,x}, \\ p_{u,v,2}(x) &= \frac{1}{(c_u - 1)(c_v - 1)} \sum_z p(z) I^{u,v}_{2,z,x}. \end{aligned} \tag{3.18}$$

**Lemma III.2** $I_{1n} = -n^{-1} \sum_x \sum_{s=1}^r a_s(x) \lambda_s + O(n^{-1}\Lambda^2 + n^{-3/2}\Lambda)$, *where* $a_s(x) = p(x) - p_{s,1}(x)$.

Proof: Noting that $O(\Lambda^2) = O(\sum_{s=1}^r \lambda_s^2) +$ terms unrelated to $\lambda$, we have

$$\begin{aligned} & E(I_{1n}) \\ &= n^{-1} \textstyle\sum_x E(L_{z,x})^2 \\ &= n^{-1} \textstyle\sum_x \sum_z p(z) (L_{z,x})^2 \\ &= n^{-1} \textstyle\sum_x \sum_z p(z) (L_{z,x})^2 [I_{0,x,z} + I_{1,x,z} + O(\Lambda^2)] \\ &= n^{-1} \{ \textstyle\sum_x p(x) [\prod_{s=1}^r (1 - \lambda_s)]^2 + \sum_x \sum_z p(z) \sum_{s=1}^r I^s_{1,x,z} \frac{\lambda_s}{c_s - 1} \prod_{u \neq s} (1 - \lambda_u) + O(\Lambda^2) \} \\ &= -n^{-1} \textstyle\sum_x \sum_{s=1}^r [p(x) - p_{s,1}(x)] \lambda_s + O(\Lambda^2)] \\ &= -n^{-1} \textstyle\sum_x a_s(x) \lambda_s + O(\Lambda^2)] \end{aligned}$$

Note that in the above $O(\Lambda^2) = O(\sum_{s=1}^r \lambda_s^2) +$ terms unrelated to $\lambda$.

Next, since $[I_{1n} - E(I_{1n})]$ has zero mean, it is easy to see that it is of order $n^{-3/2}O_p(\Lambda)$. Hence,

$I_{1n} = E(I_{1n}) + [I_{1n} - E(I_{1n})] = -n^{-1}\sum_x \sum_{s=1}^r [p(x) - p_{s,1}(x)]\lambda_s + O(n^{-1}\Lambda^2 + n^{-3/2}\Lambda)$.

**Lemma III.3** $I_{2n} = \sum_x \{\sum_{s=1}^r [p(x) - p_{s,1}(x)]\lambda_s\}^2 + O_p(n^{-1/2}\Lambda^2) + n^{-1}\sum_{s=1}^r z_{n,s}\lambda_s$, *where $z_{n,s}$ is a zero mean $O_p(1)$ random variable defined in the proof below.*

Proof: We first consider $E(I_{2n})$.

$$
\begin{aligned}
& E(L_{ij}^{(2)}) \\
&= \sum_x \sum_z \sum_w p(z)p(w)L_{z,x}L_{w,x}\{I_{0,x,z}I_{0,x,w} \\
&\quad +2I_{0,x,z}I_{1,x,w} + I_{1,x,z}I_{1,x,w} + 2I_{0,x,z}I_{2,x,w} + \Lambda^3\} \\
&= \sum_x p^2(x)\prod_{s=1}^r (1-\lambda_s)^2 + 2\sum_x p(x)\prod_{s=1}^r (1-\lambda_s)\sum_{v=1}^r p_{v,1}(x)\lambda_v \prod_{u\neq v}(1-\lambda_u) \\
&\quad + \sum_x \left[\sum_{s=1}^r p_{s,1}(x)\lambda_s \prod_{u\neq s}(1-\lambda_u)\right]^2 \\
&\quad +2\sum_x p(x)\prod_{s=1}^r (1-\lambda_s)\sum_u^r \sum_{v>u}^r p_{u,v,2}(x)\lambda_u\lambda_v \prod_{w\neq u,v}(1-\lambda_w) + \Lambda^3 \\
&= -2\sum_x p^2(x)\sum_{s=1}^r \lambda_s + 4\sum_x p^2(x)\sum_{u<v}\lambda_u\lambda_v + \sum_x p^2(x)\sum_{s=1}^r \lambda_s^2 \\
&\quad +2\sum_x p(x)\sum_{s=1}^r p_{s,1}(x)\lambda_s - 2\sum_x p(x)(\sum_{u=1}^r \lambda_u)[\sum_{s=1}^r p_{s,1}(x)\lambda_s] \\
&\quad -2\sum_x p(x)[\sum_{s=1}^r p_{s,1}(x)\lambda_s \sum_{u\neq s}\lambda_u] + \sum_x [\sum_{s=1}^r p_{s,1}(x)\lambda_s]^2 \\
&\quad +2\sum_x p(x)\sum_{1\leq u<v\leq q} p_{u,v,2}(x)\lambda_u\lambda_v + \Lambda^3.
\end{aligned}
$$

A by-product of the above result is

$$
E(L_{ij}^{(2)}) = -2\sum_x \sum_{s=1}^r p(x)[p(x) - p_{s,1}(x)]\lambda_s + O(\Lambda^2), \tag{3.19}
$$

which will be used when proving Lemma III.4 below.

$$
\begin{aligned}
& E(L_{ij}) \\
&= \sum_x \sum_z p(x)p(z)L_{z,x} \\
&= \sum_x \sum_z p(x)p(z)L_{z,x}\{I_{0,x,z} + I_{1,x,z} + I_{2,x,z} + \Lambda^3\}
\end{aligned}
$$

$$
\begin{aligned}
&= \textstyle\sum_x p^2(x) \prod_{s=1}^r (1-\lambda_s) + \sum_x \sum_{s=1}^r p(x) p_{s,1}(x) \lambda_s \prod_{u \neq s} (1-\lambda_u) \\
&\quad + \textstyle\sum_x \sum_{1 \leq u < v \leq q} p_{u,v,2}(x) \lambda_u \lambda_v \prod_{w \neq u,v} (1-\lambda_w) + \Lambda^3 \\
&= -\textstyle\sum_x p^2(x) \sum_{s=1}^r \lambda_s + \sum_x p^2(x) \sum_{u<v} \lambda_u \lambda_v \\
&\quad + \textstyle\sum_x \sum_{s=1}^r p(x) p_{s,1}(x) \lambda_s - \sum_x \sum_{s=1}^r p(x) p_{s,1}(x) \lambda_s \sum_{u \neq s} \lambda_u \\
&\quad + \textstyle\sum_x \sum_{1 \leq u < v \leq q} p(x) p_{u,v,2}(x) \lambda_u \lambda_v + \Lambda^3.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&E(L_{ij}^{(2)}) - 2E(L_{ij}) \\
&= 2 \textstyle\sum_x p^2(x) \sum_{u<v} \lambda_u \lambda_v + \sum_x p^2(x) \sum_{s=1}^r \lambda_s^2 \\
&\quad -2 \textstyle\sum_x p(x) (\sum_{u=1}^r \lambda_u) [\sum_{s=1}^r p_{s,1}(x) \lambda_s] + \sum_x [\sum_{s=1}^r p_{s,1}(x) \lambda_s]^2 + \Lambda^3 \\
&= \textstyle\sum_x [p(x) (\sum_{u=1}^r \lambda_u) - \sum_{s=1}^r p_{s,1}(x) \lambda_s]^2 + \Lambda^3 \\
&= \textstyle\sum_x \{\sum_{s=1}^r [p(x) - p_{s,1}(x)] \lambda_s\}^2 + \Lambda^3.
\end{aligned}
$$

Next, we consider $E(L_{i,j}|X_i)$.

$$
\begin{aligned}
&E(L_{i,j}|X_i) \\
&= \textstyle\sum_x p(x) L_{x_i,x} \\
&= \textstyle\sum_x p(x) L_{x_i,x} [I_{0,x_i,x} + I_{1,x_i,x} + O(\Lambda^2)] \\
&= p(X_i) \textstyle\prod_{s=1}^r (1-\lambda_s) + \sum_{s=1}^r p_{s,1}(X_i) \lambda_s \prod_{u \neq s} (1-\lambda_u) + O(\Lambda^2). \\
&= -\textstyle\sum_{s=1}^r [p(X_i) - p_{s,1}(X_i)] \lambda_s + O(\Lambda^2).
\end{aligned}
$$

Finally,

$$
\begin{aligned}
&E(L_{ij}^{(2)}|X_i) \\
&= \textstyle\sum_x \sum_z p(z) L_{x_i,x} L_{z,x} \\
&= \textstyle\sum_x \sum_z p(z) L_{x_i,x} L_{z,x} [I_{0,x_i,x} I_{0,z,x} + I_{0,x_i,x} I_{1,z,x} + I_{0,z,x} I_{1,x_i,x} + O(\Lambda^2)] \\
&= p(X_i) [\textstyle\prod_{s=1}^r (1-\lambda_s)]^2 + [\prod_{u=1}^r p_{u,1}(X_i) \lambda_u \prod_{v \neq u}^r (1-\lambda_v)] \\
&\quad + [\textstyle\prod_{s=1}^r (1-\lambda_s)] \sum_{u=1}^r p_{u,1}(X_i) \lambda_u \prod_{v \neq u}^r (1-\lambda_v)] + O(\Lambda^2)] \\
&= -2p(X_i) \textstyle\sum_{s=1}^r \lambda_s - 2[\prod_{u=1}^r p_{u,1}(X_i) \lambda_u + O(\Lambda^2)] \\
&= -2 \textstyle\sum_{s=1}^r [p(X_i) - p_{s,1}(X_i)] \lambda_s + O(\Lambda^2)].
\end{aligned}
$$

Combining the above results we obtain

$$E(L_{i,j}^{(2)}|X_i) - 2E(L_{i,j}|X_i) = 0 + (\Lambda^2). \tag{3.20}$$

Letting $K_{ij} = L_{ij}^{(2)} - 2L_{ij}$ and, by the U-statistic H-decomposition, we have
$I_{2n} = \frac{1}{n(n-1)} \sum_i \sum_{j \neq i} K_{ij} = E(K_{ij}) + 2n^{-1} \sum_{i=1}^n [E(K_{ij}|X_i) - E(K_{i,j})] + U_n,$
$= \sum_x \{\sum_{s=1}^r [p(x) - p_{s,1}(x)]\lambda_s\}^2 + O_p(n^{-1/2}\Lambda^2) + n^{-1} \sum_{s=1}^r z_{n,s}\lambda_s + O(\Lambda^2),$
where $U_n = 2(n(n-1))^{-1} \sum_{i=1}^n \sum_{j>i}^n [K_{ij} - E(K_{ij}|X_i) - E(K_{ij}|X_j) + E(K_{ij})]$ is a degenerate U-statistic. Obviously, $U_n$ has zero mean and is of order $O_p(n^{-1})$. $n^{-1} \sum_{s=1}^r z_{n,s}\lambda_s$ represent terms in $U_n$ that are linear in $\bar{\lambda}$. A proof for this is similar to the proof of Lemma III.11 and is thus omitted here.

**Lemma III.4** *$I_{3n} = -n^{-1} \sum_{s=1}^r b_s(x)\lambda_s + O_p(n^{-3/2}) +$ terms unrelated to $\lambda$, where $b_s(x) = -2 \sum_x p(x)[p(x) - p_{s,1}(x)]$.*

Proof: By (3.19) we have $E(L_{ij}^{(2)}) = -\sum_{s=1}^r b_s(x)\lambda_s + O_p(\Lambda^2)$, and using this result and the U-statistic H-decomposition, we have

$I_{3n} = \frac{1}{n^2(n-1)} \sum_i \sum_{j \neq i} L_{i,j}^{(2)}$
$= n^{-1}\{E(L_{ij}^{(2)}) + 2n^{-1} \sum_i [E(L_{ij}^{(2)}|X_i) - E(L_{ij}^{(2)})] + O_p(n^{-1})\}$
$= -n^{-1}[\sum_{s=1}^r b_s(x)\lambda_s] + O_p(n^{-3/2} + n^{-1}\lambda^2) +$ terms unrelated to $\lambda$.

**Lemma III.5** *Define an $r \times r$ matrix $\Omega$ by $\Omega_{st} = \sum_x [p(x) - p_{s,1}(x)][p(x) - p_{t,1}(x)]$. Then $\Omega$ is positive-definite.*

Proof: Letting $\lambda'\Omega\lambda = \sum_x \{\sum_{s=1}^r [p(x) - p_{s,1}(x)]\lambda_s\}^2 = 0$, then for any $x$, we must have

$$\sum_{s=1}^r [p(x) - p_{s,1}(x)]\lambda_s = 0.$$

Summing the above equation over $x^2, \ldots, x^r$, and noting that $\sum_{x^2,\ldots,x^r}[p(x) - p_{s,1}(x)] = 0$ for $s = 2, \ldots, r$, for all $x^1 \in \{0, 1, \ldots, c_1 - 1\}$ we get

$$[p_1(x^1) - p_{1,1}(x^1)]\lambda_1 = 0, \tag{3.21}$$

where $p_1(x^1) = \sum_{x^2,\ldots,x^r} p(x)$ is the marginal probability of $x^1$, and $p_{1,1}(x^1) = \sum_{x^2,\ldots,x^r} p_{s,1}(x)$. Since $p_1(x^1)$ can not be a constant for all $x^1$, we know that there exists $x^1 \in \{0, 1, \ldots, c_1 - 1\}$ for which $p_1(x^1) - p_{1,1}(x^1) \neq 0$, for otherwise $x^1$ would be a uniform component, which contradicts our assumption. Therefore, we must have $\lambda_1 = 0$.

By symmetry we know that we must have $\lambda_s = 0$ for all $s = 1, \ldots, r$. This completes the proof that $\Omega$ is positive definite.

**Proof of Equation (3.9)**

Summarizing lemmas III.2 to III.4, we have shown that

$$\begin{aligned}
&CV(\lambda) = I_{1n} + I_{2n} - I_{3n} \\
&= \textstyle\sum_x \{\sum_{s=1}^r [p(x) - p_{s,1}(x)]\lambda_s\}^2 - n^{-1} \sum_x \sum_{s=1}^r [a_s(x) + b_s(x) + z_{n,s}]\lambda_s + (s.o.) \\
&= \textstyle\sum_x \{\sum_{s=1}^r [p(x) - p_{s,1}(x)]\lambda_s\}^2 - n^{-1} \sum_{s=1}^r c_{n,s}\lambda_s + (s.o.) \\
&\equiv \lambda'\Omega\lambda - n^{-1}\lambda' c_n + (s.o.),
\end{aligned}$$

where $c_n = (c_{n1}, \ldots, c_{nr})'$ with $c_{ns} = a_s(x) + b_s(x) + z_{ns}$, and $(s.o.) = o(n^{-1}\Lambda + \Lambda^2)$ denote terms of smaller order or terms unrelated to $\lambda$.

**Proof of Theorem III.3**

Under the conditions of Theorem III.3, the proofs of lemmas III.1 through Lemma III.4 are similar to the previous derivation. Therefore, here we only prove Lemma III.5 under the conditions of Theorem III.3.

Similar to the proof of Lemma III.5, we know that if $\lambda'\Omega\lambda = 0$, then

$$\sum_{s=1}^{r}[p(x) - p_{s,1}(x)]\lambda_s = 0$$

for all $x \in \mathcal{S}$.

Summing the above equation over $x^3, \ldots, x^r$, and noting that $\sum_{x^3,\ldots,x^r}[p(x) - p_{s,1}(x)] = 0$ for $s = 3, \ldots, r$, for all $x^1, x^2 \in \mathcal{S}_1 \times \mathcal{S}_2$ we obtain

$$\begin{aligned} 0 &= \left[\frac{c_1}{c_1 - 1}p_{j,1,2}(x^1, x^2) - \frac{1}{c_1 - 1}p_2(x^2)\right]\lambda_1 \\ &\quad + \left[\frac{c_2}{c_2 - 1}p_{j,1,2}(x^1, x^2) - \frac{1}{c_2 - 1}p_1(x^1)\right]\lambda_2 \\ &= \left[\frac{c_1}{c_1 - 1}\lambda_1 + \frac{c_2}{c_2 - 1}\lambda_2\right]\left[p(x^1, x^2) - \frac{1}{c_1 c_2}\right] \end{aligned} \tag{3.22}$$

for all $x^1$, $x^2$, where $p_{j,1,2}(x^1, x^2)$ is the joint probability of $(x^1, x^2)$, and we have also used the condition of Theorem III.3 that $p_j(x^j) = 1/c_j$ $(j = 1, 2)$.

Since $p_{j,1,2}(x^1, x^2) \neq 1/(c_1 c_2)$ for at least some $x^1, x^2$, we must have $\lambda_1 = \lambda_2 = 0$. By Lemma III.5 we know that we also must have $\lambda_s = 0$ for $s = 3, \ldots, r$. Therefore, $\Omega$ is positive definite under the conditions of Theorem III.3.

## 2. Some Useful Lemmas

We use $\bar{x}$ and $\bar{\lambda}$ to denote non-uniformly distributed variables and their smoothing parameters, and use $\tilde{x}$ and $\tilde{\lambda}$ for the uniformly distributed variable and its smoothing parameter. We write $L_{x_i,x} = L_{\bar{x}_i,\bar{x}}L_{\tilde{x}_i,\tilde{x}}$, and let $I_{jn}$ be defined as in subsection F.1. We first present some lemmas.

In Lemma III.7 below we show that when there exists a uniformly distributed variable $(\tilde{x})$, the leading terms of $CV(.)$, $E(I_{jn})$ $(j = 1, 2, 3)$, are all unrelated to $\tilde{\lambda}$. Therefore, by exactly the same proofs as given in subsection F.1. one can show that $\hat{\lambda}_s = O_p(n^{-1})$ for $s = 1, \ldots, r-1$. Therefore, for the remainder of subsection F.2. we

need only prove Theorem III.4 (ii).

**Lemma III.6** $I_{1n} = n^{-1}A_n(\bar{\lambda})B_n(\tilde{\lambda}) + O_p(n^{-3/2})$,
*where* $A_n(\bar{\lambda}) = \sum_{\bar{x}}\sum_{\bar{z}}\bar{p}(\bar{z})(L_{\bar{z},\bar{x}})^2$, *and* $B_n(\tilde{\lambda}) = [(1-\tilde{\lambda})^2 + (c-1)(\frac{\tilde{\lambda}}{c-1})^2]$ $(c \equiv c_r)$.

Proof: $I_{1n} = n^{-2}\sum_i L_{ii}^{(2)} = n^{-2}\sum_{i=1}^{n}\sum_x (L_{x_i,x})^2$, and noting that $p(z) = \bar{p}(\bar{z})/c$ is unrelated to $\tilde{z}$, we have

$$E(I_{1n}) = n^{-1}\sum_x\sum_z p(z)(L_{z,x})^2 = (nc)^{-1}\sum_{\bar{x}}\sum_{\bar{z}}\bar{p}(\bar{z})(L_{\bar{z},\bar{x}})^2\sum_{\tilde{x}}\sum_{\tilde{z}}(L_{\tilde{z},\tilde{x}})^2$$
$$= (nc)^{-1}\sum_{\bar{x}}\sum_{\bar{z}}\bar{p}(\bar{z})(L_{\bar{z},\bar{x}})^2\sum_{\tilde{x}}\sum_{\tilde{z}}[(1-\tilde{\lambda})^{2I_{\tilde{z}=\tilde{z}}}(\tfrac{\tilde{\lambda}}{c-1})^{2I_{\tilde{z}\neq\tilde{x}}}]$$
$$= n^{-1}\sum_{\bar{x}}\sum_{\bar{z}}\bar{p}(\bar{z})(L_{\bar{z},\bar{x}})^2[(1-\tilde{\lambda})^2 + (c-1)(\tfrac{\tilde{\lambda}}{c-1})^2]$$
$$= n^{-1}A_n(\bar{\lambda})B_n(\tilde{\lambda}).$$

From $I_{1n} - E(I_{1n})$ having zero mean it is easy to see that $I_{1n} - E(I_{1n}) = O_p(n^{-3/2})$. Therefore, we have

$$I_{1n} = E(I_{1n}) + (I_{1n} - E(I_{1n})) = E(I_{1n}) + O_p(n^{-3/2}) = n^{-1}A_n(\bar{\lambda})B_n(\tilde{\lambda}) + O_p(n^{-3/2}).$$

**Lemma III.7** $E(L_{ij})$, $E(L_{ij}^{(2)})$, $E(L_{ij}|X_i)$, $E(L_{ij}^{(2)}|X_i)$ *are all unrelated to* $\tilde{\lambda}$.

Proof: We first compute $E(L_{ij}^{(2)})$.

$$E(L_{ij}^{(2)}) = \sum_x\sum_z\sum_w p(z)p(w)L_{z,x}L_{w,x} = \sum_x[\sum_z p(z)L_{z,x}]^2$$
$$= \tfrac{1}{c^2}\sum_x[\sum_{\bar{z}}\bar{p}(\bar{z})L_{\bar{z},\bar{x}}\sum_{\tilde{z}}L_{\tilde{z},\tilde{x}}]^2 = \tfrac{1}{c^2}\sum_x[\sum_{\bar{z}}\bar{p}(\bar{z})L_{\bar{z},\bar{x}}]^2$$

because $\sum_{\tilde{z}} l(\tilde{z},\tilde{x},\tilde{\lambda}) = 1$. The reason why $E(L_{ij}^{(2)})$ is unrelated to $\tilde{\lambda}$ is because $p(z) = \bar{p}(\bar{z})/c$, which is unrelated to $\tilde{z}$.

Similarly (again using $\sum_{\tilde{z}} l(\tilde{z},\tilde{x},\tilde{\lambda}) = 1$),

$E(L_{ij}) = \sum_x p(x)\sum_z p(z)L_{z,x} = c^{-1}\sum_{\bar{x}}\bar{p}(\bar{x})\sum_{\bar{z}}\bar{p}(\bar{z})L_{\bar{z},\bar{x}}$ which is unrelated to $\tilde{\lambda}$,

where we used $p(x) = \bar{p}(\bar{x})/c$ and $\sum_{\tilde{x}}\sum_{\tilde{z}}L_{\tilde{z},\tilde{x}} = \sum_{\tilde{x}}1 = c$. Similarly,

$E(L_{ij}|X_i) = \sum_x p(x)L_{x,x_i} = c^{-1}\sum_{\bar{x}}\bar{p}(\bar{x})L_{\bar{x}_i,\bar{x}}$ which is also unrelated to $\tilde{\lambda}$.

**Lemma III.8** *$I_{3n} = O_p(n^{-2}) +$ terms unrelated to $\tilde{\lambda}$.*

Proof: By the U-statistic H-decomposition, we have

$$\begin{aligned} & I_{3n} \\ &= n^{-1}[\tfrac{2}{n(n-1)} \textstyle\sum_i \sum_{j>i} L_{i,j}^{(2)}] \\ &= n^{-1}\{E(L_{ij}^{(2)}) + \tfrac{2}{n} \textstyle\sum_i [E(L_{ij}^{(2)}|X_i) - E(L_{ij}^{(2)})] + O_p(n^{-1})\} \\ &= O_p(n^{-2}) + \text{ terms unrelated to } \tilde{\lambda} \end{aligned}$$

because $E(L_{ij}^{(2)})$ and $E(L_{ij}^{(2)}|X_i)$ are unrelated to $\tilde{\lambda}$ by Lemma III.7.

**Lemma III.9** *Define $z_{n,1} = \frac{2}{n-1} \sum_i \sum_{j>i} H_{ij}$, where $H_{ij} = I_{\bar{x}_i=\bar{x}_j}[\frac{1}{c-1} I_{\tilde{x}_i \neq \tilde{x}_j} - I_{\tilde{x}_i=\tilde{x}_j}]$. Then $n^{-1}z_{n,1}$ is a degenerate U-statistic of order $O_p(n^{-1})$.*

Proof: Letting $\bar{I}_{ij} = I_{\bar{x}_i=\bar{x}_j}$ and noting that $p(x) = \bar{p}(\bar{x})/c$, we have

$$\begin{aligned} & E[H_{ij}|X_i] = \tfrac{1}{c-1} E[\bar{I}_{ij} I_{\tilde{x}_i \neq \tilde{x}_j}|X_i] - E[\bar{I}_{ij} I_{\tilde{x}_i=\tilde{x}_j}|X_i] \\ &= \tfrac{1}{c(c-1)} \bar{p}(\bar{X}_i) \textstyle\sum_{\tilde{z}} I_{\tilde{z} \neq \tilde{x}_i} - \bar{p}(\bar{X}_i)/c \\ &= \bar{p}(\bar{X}_i)/c - \bar{p}(\bar{X}_i)/c = 0 \text{ (because } \textstyle\sum_{\tilde{z}} I_{\tilde{z} \neq \tilde{x}_i} = c-1). \end{aligned}$$

Note that $E[H_{ij}|X_i] = 0$ implies that $E[H_{ij}] = 0$. Therefore, $n^{-1}z_{n,1} = \frac{2}{n(n-1)} \sum_i \sum_{j>i} H_{ij}$ is a degenerate U-statistic of order $O_p(n^{-1})$ (since $E[(n^{-1}z_{n,1})^2] = O(n^{-2})$), which implies that $z_{n,1}$ is a zero mean $O_p(1)$ random variable.

**Lemma III.10** *Let $A_{1,ij} = \sum_{\tilde{x}} I_{\tilde{x}_i=\tilde{x}} I_{\tilde{x}_j=\tilde{x}}$, $A_{2,ij} = \frac{1}{c-1} \sum_{\tilde{x}} [I_{\tilde{x}_i \neq \tilde{x}} I_{\tilde{x}_j=\tilde{x}} + I_{\tilde{x}_i=\tilde{x}} I_{\tilde{x}_j \neq \tilde{x}}]$, let $A_{3,ij} = \frac{1}{(c-1)^2} \sum_{\tilde{x}} \tilde{I}_{x_i \neq x} \tilde{I}_{x_j \neq x}$, and define (note that $\bar{I}_{ij} = I_{\bar{x}_i=\bar{x}_j}$)*

$z_{n,2} = \frac{2}{n-1} \sum_i \sum_{j>i} \bar{I}_{ij}[A_{2,ij} - 2A_{1,ij}]$, *and*

$z_{n,3} = \frac{2}{n-1} \sum_i \sum_{j>i} \bar{I}_{ij}[A_{1,ij} - A_{2,ij} + A_{3,ij}]$.

*Then both $n^{-1}z_{n,2}$ and $n^{-1}z_{n,3}$ are degenerate U-statistics of order $O_p(n^{-1})$.*

Proof: The proof is similar to the proof for Lemma III.9. Here we only provide a proof for $z_{n,2}$. It suffices to show that $E[\bar{I}_{ij}(A_{2,ij} - 2A_{1,ij})|X_i] = 0$.

$E[\bar{I}_{ij}(A_{2,ij} - 2A_{1,ij})|X_i] = \bar{p}(\bar{X}_i)\frac{1}{c(c-1)}\sum_{\tilde{x}}\sum_{\tilde{z}}[I_{\tilde{X}_i\neq\tilde{x}}I_{\tilde{z}=\tilde{x}} + I_{\tilde{X}_i=\tilde{x}}I_{\tilde{z}\neq\tilde{x}}] - 2\bar{p}(\bar{X}_i)/c$
$= \bar{p}(\bar{X}_i)\frac{1}{c(c-1)}[(c-1)+(c-1)] - 2\bar{p}(\bar{X}_i)(1/c) = 0.$

**Lemma III.11** *$I_{2n} = n^{-1}\mathcal{Z}_{n,1}\tilde{\lambda} + \frac{1}{2}n^{-1}\mathcal{Z}_{n,2}\tilde{\lambda}^2 + O_p(n^{-3/2}) +$ terms unrelated to $\tilde{\lambda}$, where $\mathcal{Z}_{n,1}$ and $\mathcal{Z}_{n,2}$ are both zero mean $O_p(1)$ random variables.*

Proof: $I_{2n} = \frac{2}{n(n-1)}\sum_i\sum_{j>i}[L_{ij}^{(2)} - 2L_{ij}]$. Since $\bar{L}_{ij} = \bar{I}_{ij} + O_p(n^{-1})$ we will replace $\bar{L}_{ij}$ by $\bar{I}_{ij} \equiv I_{\bar{x}_i=\bar{x}_j}$ and put the $O_p(n^{-1})$ term in (s.o.) (as it is of smaller order and is unrelated to $\tilde{\lambda}$). Similarly, we replace $\bar{L}_{\bar{x}_i,\bar{x}}$ and $\bar{L}_{\bar{x}_j,\bar{x}}$ by $I_{\bar{x}_i=\bar{x}}$ and $I_{\bar{x}_j=\bar{x}}$.

Thus, we have $L_{ij} = \bar{I}_{\bar{x}_i=\bar{x}_j}\tilde{L}_{ij} + (s.o.) = \bar{I}_{\bar{x}_i=\bar{x}_j}\tilde{L}_{ij}[I_{\tilde{x}_i=\tilde{x}_j} + I_{\tilde{x}_i\neq\tilde{x}_j}] + (s.o.)$. Noting that $\tilde{L}_{ij}I_{\tilde{x}_i=\tilde{x}_j} = (1-\tilde{\lambda})I_{\tilde{x}_i=\tilde{x}_j}$, and $\tilde{L}_{ij}I_{\tilde{x}_i\neq\tilde{x}_j} = (\tilde{\lambda}/(c-1))I_{\tilde{x}_i\neq\tilde{x}_j}$, we have

$$\begin{aligned}
& \frac{2}{n(n-1)}\sum_i\sum_{j>i} L_{ij} \\
= \; & \frac{2}{n(n-1)}\sum_i\sum_{j>i}\bar{I}_{ij}\tilde{L}_{ij}\left[I_{\tilde{x}_i=\tilde{x}_j} + I_{\tilde{x}_i\neq\tilde{x}_j}\right] + (s.o.) \\
= \; & \frac{2}{n(n-1)}\sum_i\sum_{j>i}\bar{I}_{ij}\left\{I_{\tilde{x}_i=\tilde{x}_j} + \tilde{\lambda}\left[\frac{1}{c-1}I_{\tilde{x}_i=\tilde{x}_j} - I_{\tilde{x}_i=\tilde{x}_j}\right]\right\} \\
= \; & n^{-1}\tilde{\lambda}z_{n,1} + \text{ terms unrelated to } \tilde{\lambda}, \qquad (3.23)
\end{aligned}$$

where $z_{n,1} = \frac{2}{(n-1)}\sum_i\sum_{j>i}H_{ij}$ with $H_{ij} = \bar{I}_{ij}[\frac{1}{c-1}I_{\tilde{x}_i\neq\tilde{x}_j} - I_{\tilde{x}_i=\tilde{x}_j}]$. By Lemma III.9 we know that $z_{n,1}$ is a zero mean $O_p(1)$ random variable.

Next we consider $L_{ij}^{(2)}$.

$$\begin{aligned}
L_{ij}^{(2)} &= \sum_{\bar{x}} \bar{I}_{ix}\bar{I}_{jx} \sum_{\tilde{x}} \tilde{L}_{ix}\tilde{L}_{jx} + (s.o.) \\
&= \bar{I}_{ij} \sum_{\tilde{x}} \tilde{L}_{ix}\tilde{L}_{jx} = \bar{I}_{ij} \sum_{\tilde{x}} \tilde{L}_{ix}\tilde{L}_{jx}[\tilde{I}_{0,ix} + \tilde{I}_{1,ix}][\tilde{I}_{0,jx} + \tilde{I}_{1,jx}] + (s.o.) \\
&= \bar{I}_{ij} \sum_{\tilde{x}} \tilde{L}_{ix}\tilde{L}_{jx}[\tilde{I}_{0,ix}\tilde{I}_{0,jx} + \tilde{I}_{1,ix}\tilde{I}_{0,jx} + \tilde{I}_{0,ix}\tilde{I}_{1,jx} + \tilde{I}_{1,ix}\tilde{I}_{1,jx}] + (s.o.) \\
&= \bar{I}_{ij}\{(1-\tilde{\lambda})^2 \sum_{\tilde{x}} I_{\tilde{x}_i=\tilde{x}} I_{\tilde{x}_j=\tilde{x}} + \sum_{\tilde{x}} \frac{(1-\tilde{\lambda})\tilde{\lambda}}{c-1}[I_{\tilde{x}_i\neq\tilde{x}} I_{\tilde{x}_j=\tilde{x}} + I_{\tilde{x}_i=\tilde{x}} I_{\tilde{x}_j\neq\tilde{x}}] \\
&\quad + \sum_{\tilde{x}} \frac{\tilde{\lambda}^2}{(c-1)^2} \tilde{I}_{X_i\neq x} I_{\tilde{x}_j\neq\tilde{x}}\} + (s.o.) \\
&= \bar{I}_{ij}\{A_{1,ij} + \tilde{\lambda}[-2A_{1,ij} + A_{2,ij}] + \tilde{\lambda}^2[A_{1,ij} - A_{2,ij} + A_{3,ij}]\} + (s.o.)
\end{aligned}$$

where $A_{s,ij}$ $(s = 1, 2, 3)$ are the same as defined in Lemma III.10.

Therefore,

$$\begin{aligned}
&\frac{2}{n(n-1)} \sum_i \sum_{j>i} L_{ij}^{(2)} \\
= &\frac{2}{n(n-1)} \sum_i \sum_{j>i} \bar{I}_{ij} \Big\{ A_{1,ij} + \tilde{\lambda}[A_{2,ij} - 2A_{1,ij}] \\
&+\tilde{\lambda}^2[A_{1,ij} - A_{2,ij} + A_{3,ij}] \Big\} + (s.o.) \\
= &\; n^{-1}\tilde{\lambda} z_{n,2} + n^{-1}\tilde{\lambda}^2 z_{n,3} + (s.o.) + \text{ terms related to } \tilde{\lambda}, \qquad (3.24)
\end{aligned}$$

where $z_{n,2} = \frac{2}{n(n-1)} \sum_i \sum_{j>i} \bar{I}_{ij}[A_{2,ij} - 2A_{1,ij}]$ and $z_{n,3} = \frac{2}{n(n-1)} \sum_i \sum_{j>i} \bar{I}_{ij}[A_{1,ij} - A_{2,ij} + A_{3,ij}]$.

By Lemma III.10 we know that $z_{n,2}$ and $z_{n,3}$ are both zero mean $O_p(1)$ random variables.

Combining (3.23) and (3.24) we have

$$U_n = n^{-1}\tilde{\lambda}\mathcal{Z}_{n,1} + \frac{1}{2}n^{-1}\tilde{\lambda}^2\mathcal{Z}_{n,2} + (s.o.) + \text{ terms related to } \tilde{\lambda}, \qquad (3.25)$$

where $\mathcal{Z}_{n,1} = z_{n,2} - 2z_{n,1}$ and $\mathcal{Z}_{n,2} = 2z_{n,3}$ are both zero mean $O_p(1)$ random variables. This completes the proof of Lemma III.11.

**Lemma III.12** *Asymptotically, there is a positive probability ($\delta_1 \in (0,1)$) that $U_n$ is minimized at $\tilde{\lambda} = (c-1)/c$ with probability $\delta_1$, and with probability $1-\delta_1$ that $U_n$ is minimized for $\tilde{\lambda} \in [0,(c-1)/c)$ $(c \equiv c_r)$.*

Proof: Taking the derivative of $U_n$ with respect to $\tilde{\lambda}$ we get

$$U^{(1)}_{n,\tilde{\lambda}} \stackrel{def}{=} \frac{\partial U_n}{\partial \tilde{\lambda}} = n^{-1}[\mathcal{Z}_{n,1} + \tilde{\lambda}\mathcal{Z}_{n,2}]. \tag{3.26}$$

It is easy to see that if

$$\mathcal{Z}_{n,1} + \tilde{\lambda}\mathcal{Z}_{n,2} < 0 \text{ for all } \tilde{\lambda} \in [0,(c-1)/c] \ , \tag{3.27}$$

then $U_n$ is minimized at $\tilde{\lambda} = (c-1)/c$. Obviously (3.27) occurs with positive probability. This is because both $\mathcal{Z}_{n,1}$ and $\mathcal{Z}_{n,2}$ are zero mean $O_p(1)$ random variables. For example, if $\mathcal{Z}_{n,1} < 0$ and $\mathcal{Z}_{n,2} \leq 0$, then (3.27) holds true; or if $\mathcal{Z}_{n,1} < 0$ and $\mathcal{Z}_{n,2} > 0$, but with $\mathcal{Z}_{n,1} + \frac{c-1}{c}\mathcal{Z}_{n,2} < 0$, (3.27) also holds.

When $n \to \infty$, $\mathcal{Z}_{n,j}$ converges to a well defined zero mean $O_p(1)$ random variable, say $\mathcal{Z}_{j,\infty}$ $(j = 1,2)$. Therefore, asymptotically, $U_n$ is minimized at $\tilde{\lambda} = (c-1)/c$ with a constant probability $\delta_1 \in (0,1)$.

**Proof of Theorem III.4**

Summarizing lemmas III.6 through III.8 above we have shown that

$$\begin{aligned} CV(\lambda) &= I_{1n} + I_{2n} + I_{3n} \\ &= n^{-1}A_n(\bar{\lambda})B_n(\tilde{\lambda}) + \tilde{\lambda}\mathcal{Z}_{n,1} + \frac{1}{2}\tilde{\lambda}^2\mathcal{Z}_{n,2} + (s.o.), \\ &= n^{-1}B_n(\tilde{\lambda}) + \tilde{\lambda}\mathcal{Z}_{n,1} + \frac{1}{2}\tilde{\lambda}^2\mathcal{Z}_{n,2} + (s.o.) \end{aligned} \tag{3.28}$$

where the last equality used $A_n(\bar{\lambda}) = 1 + O_p(n^{-1})$ because $\bar{\lambda} = O_p(n^{-1})$, which leads to $A_n(\bar{\lambda}) = 1 + O_p(n^{-1})$.

The first term that is related to $\tilde{\lambda}$ in (3.28) is (except for the $n^{-1}$ factor and

omitting smaller order terms) $B_n(\tilde{\lambda}) = (1-\tilde{\lambda})^2 + (c-1)(\frac{\tilde{\lambda}}{c-1})^2$.

To minimize this term, $\tilde{\lambda}$ must satisfy

$$B^{(1)}_{n,\tilde{\lambda}} = \frac{\partial B_n(\tilde{\lambda})}{\partial \lambda_k} = 2\left[-(1-\tilde{\lambda}) + (\frac{\tilde{\lambda}}{c-1})\right] = 0, \tag{3.29}$$

and solving for $\tilde{\lambda}$ gives $\tilde{\lambda} = \frac{c-1}{c}$. Thus,

$$B_n(\tilde{\lambda}) \text{ is minimized when } \tilde{\lambda} \text{ assumes its upper extreme value.} \tag{3.30}$$

The second leading term that is related to $\tilde{\lambda}$ is $\tilde{\lambda}\mathcal{Z}_{n,1} + (1/2)\tilde{\lambda}^2\mathcal{Z}_{n,2}$. Its derivative with respect to $\tilde{\lambda}$ is

$$\mathcal{Z}_{n,1} + \tilde{\lambda}\mathcal{Z}_{n,2}. \tag{3.31}$$

By Lemma III.12 we know that (3.31) is minimized at $\tilde{\lambda} = (c-1)/c$ with a positive probability $\delta_1$. (3.29) to (3.31) imply that $CV(\tilde{\lambda})$ is minimized at $\tilde{\lambda} = (c-1)/c$ with a positive probability $\delta > \delta_1$ ($\delta \in (0,1)$).

# CHAPTER IV

# K-NN ESTIMATION OF ECONOMETRIC MODELS

## A. Introduction

It is well known that nonparametric estimation techniques have the advantage of being robust to functional form specifications. Nonparametric kernel method, series method and k nearest neighbor (k-nn) method are all widely used by applied econometricians. It is also well established that the selection of smoothing parameters are of crucial importance in nonparametric estimations. In nonparametric regression model estimation based on kernel or series techniques, various data-driven methods are developed. With independent data, the least squared leave-one-out cross-validation method based on the kernel or series estimators is the most popular choice of selecting the smoothing parameters. Intuitively, one can also apply the leave-one-out cross-validation method to selecting the smoothing parameter when using the k-nn nonparametric estimation method. However, to the best of out knowledge, the asymptotic behavior of the cross-validation selected smoothing parameter in a k-nn framework is not available in the literature. In this chapter we consider both a local constant and a local linear k-nn estimator and we provide asymptotic analysis for the cross-validation selected $k$ in both local constant and local linear k-nn regression function estimations.

Different nonparametric estimation methods have their own advantages in different situations. Undoubtedly k-nn method might be the preferred method in some applications. For example, when data points are unevenly distributed in its support, as is often the case for nonexperimental data in social sciences, kernel method with a fixed smoothing parameter (over the whole data range) may contain few data points in certain range of the support and thus may lead to unreliable estimation and infer-

ence results. In this case a researcher may prefer using a k-nn method which always uses (the nearest) $k$ data points in the nonparametric estimation. Although we will only consider the independent data case in this chapter, the results of the chapter can be extended to the weakly dependent case (say $\beta$ or $\alpha$ mixing processes) in a straightforward way.

The chapter is organized as follows. In section B we consider using the least squares cross-validation method to select the smoothing parameter in the k-nn local constant estimation. We show that the cross-validation selected smoothing parameter converges to a non-stochastic optimal value which minimizes the asymptotic a weighted estimation mean square errors. Section C considers the local linear k-nn estimation case. All the proofs are given in the section D.

## B. Cross-Validation with Local Constant k-nn Estimation

We consider the nonparametric regression model

$$Y_i = g(X_i) + u_i, \qquad i = 1, 2, ..., n, \tag{4.1}$$

where $X_i \in R^p$ is a vector of dimension $p$, the functional form of $g(.)$ is unspecified. We only consider the case that $(Y_i, X_i)_{i=1}^n$ are independent and identically distributed (i.i.d), though the results of the chapter can be readily extended to weakly dependent time series data case.

For a fixed value $x \in R^p$, define the k-nearest-neighbor (k-nn) distance, centered as $x$ by

$$\begin{aligned} &R_n \equiv R_n(x) \stackrel{def}{=} \text{the } k\text{th-nearest-neighbor Euclidean distance to } x \\ &\text{among all the } X_j\text{'s, for } j = 1, ..., n. \end{aligned} \tag{4.2}$$

We also define the $k$-nn distance centered at $X_i$ as:

$$R_i \equiv R_n(X_i) \stackrel{def}{=} \text{the } k\text{th-nearest-neighbor Euclidean distance to } X_i \text{ among all the } X_j\text{'s, for } j \neq i. \tag{4.3}$$

In this section we consider k-nn local constant estimator of $g(x)$. Let $w(\cdot) : R^p \to R$ be a bounded non-negative weight function, $w(v) = w(-v)$, $\int w(v)dv = 1$, $w(v) = 0$ for $||v|| \geq 1$, where $||v||$ denotes the Euclidean norm of $v$, the local constant k-nn estimator of $g(x)$ is given by

$$\hat{g}(x) = \frac{1}{nR_n^p} \sum_{i=1}^{n} Y_i \, w\left(\frac{X_i - x}{R_n}\right) / \hat{f}(x) \tag{4.4}$$

where

$$\hat{f}(x) = \frac{1}{nR_n^p} \sum_{i=1}^{n} w\left(\frac{X_i - x}{R_n}\right) \tag{4.5}$$

is the k-nn estimator of $f(x)$.

The main difference between a kernel estimator and the k-nn estimator $g(x)$ is that the bandwidth $R_n$ in the k-nn estimation is stochastic. Hence the asymptotic analysis of cross-validation method with k-nn estimation is more complex than that with kernel estimation.

We will choose $k$ by minimizing the following cross-validation function

$$CV(k) = n^{-1} \sum_{i=1}^{n} (Y_i - \hat{g}_{-i}(X_i))^2 \, M(X_i), \tag{4.6}$$

where $\hat{g}_{-i}(X_i) = \sum_{j \neq i}^{n} Y_j w\left(\frac{X_j - X_i}{R_i}\right) / \sum_{j \neq i}^{n} w\left(\frac{X_j - X_i}{R_i}\right)$ is the leave-one-out k-nn estimator of $g(X_i)$, and $M(.)$ is a weight function that avoids the too small a value for the denominator of $CV(k)$. We use $\hat{k}$ to denote the cross-validation selected value of $k$. The following assumptions will be used in studying the asymptotic behavior of $\hat{k}$.

**Assumption 1**: $(X_i, Y_i)$ are independent and identically distributed, $u_i = Y_i - g(X_i)$,

$E(u_i^4)$ is finite. $g(x)$ and $f(x)$ are both continuous differentiable up to the third order. $\sigma^2(x) = E(u_i^2|X_i = x)$ is continuous in $x$. Let $\mathcal{S}$ denote the support of $M(.)$, then $inf_{x\in\mathcal{S}} f(x) \geq \delta$ for some $\delta > 0$.

**Assumption 2**: $w(.)$ is a bounded symmetric non-negative function, $w(v) = 0$ for $||v|| \geq 1$, $\int w(v)dv = \int_{||v||\leq 1} w(v)dv = 1$, $\int w(v)vv'dv = c_w I$, $\int w^2(v)dv = d_w$ and $\int w^2(v)vv'du = \nu_w I$, where $I$ is a identity matrix of dimension $p$. $c_w$, $d_w$ and $\nu_w$ are all finite positive constants.

**Assumption 3**: $k \in \Lambda = [n^\epsilon, n^{1-\epsilon}]$ for some arbitrarily small $\epsilon \in (0, 1/2)$.

Assumption 1 contains some standard smoothness and moment conditions on $g(.)$ and $f(.)$. Assumption 2 states that $w(.)$ is a second order kernel function which vanishes at the boundary and outside the unit sphere. The condition that $w(v) = 0$ for $||v|| \geq 1$ can be relaxed to $w(v) = 0$ for $||v|| > 1$ without changing the results of the chapter, but it needs longer proofs. Assumption 3 implies that $k/n \to 0$ and $k \to \infty$ as $n \to \infty$, which ensures that the bias and variance of k-nn estimation converge to zero as sample size increases.

Denote the $p \times 1$ first derivative vector and the $p \times p$ second derivative matrix of $g(x)$ by $\nabla g(x)$ and $\nabla^2 g(x)$, respectively. Also let $g_s(x)$ ($f_s(x)$) denote the partial derivative of $g$ ($f$) with respect to the $s$th component of $x$ ($s = 1, ..., p$).

In the D.1. we show that, uniformly in $k \in \Lambda$, the leading term of $CV(k)$ has the form of $\Phi_1(k/n)^{4/p} + \Phi_2 k^{-1}$, where

$$\Phi_1 = c_0^{-4/p} c_w^2 \int \left( \left[ \frac{1}{2} f(x)\, tr[\nabla^2 g(x) + \sum_{s=1}^{p} f_s(x)\, g_s(x) \right] \right)^2 [f(x)]^{-(p+4)/p} M(x)\, dx, \tag{4.7}$$

$c_0 = \pi^{p/2}/\Gamma((p+2)/2)$ is the volume of unit ball in $R^p$, and

$$\Phi_2 = c_0 d_w \int \sigma^2(x)\, M(x)\, f(x)\, dx. \tag{4.8}$$

Thus, we have

$$CV(k) = \Phi_1(k/n)^{4/p} + \Phi_2 k^{-1} + o((k/n)^{4/p} + k^{-1}) \tag{4.9}$$

uniformly in $k \in \Lambda$. Let $k_0$ denote the value of $k$ that minimizes $\Phi_1(k/n)^{4/p} + \Phi_2 k^{-1}$, the leading term of $CV(k)$. Then it is easy to show that $k_0 = a_0 n^{4/(4+p)}$, where $a_0 = [p\Phi_2/(4\Phi_1)]^{p/(4+p)}$. Recall that $\hat{k}$ is the cross-validation selected value of $k$, then from (4.9) we immediately have the following result.

**Theorem IV.1** *Under assumptions 1 to 3, we have*

$\hat{k} = k_0 + o_p(k_0)$ *or equivalently* $\hat{k}/k_0 \xrightarrow{p} 1$.

The proof of Theorem IV.1 is given in the D.1.

In the next section, we study cross-validation selection of $k$ in k-nn local linear estimation.

C. Cross-Validation with Local Linear k-nn Estimation

The nonparametric regression model is the same as we considered in section B

$$Y_i = g(X_i) + u_i, \qquad (i = 1, ..., n) \tag{4.10}$$

Denotes by $\delta(x) = (g(x), (\nabla g(x))')'$. Note that $\delta(x)$ is of dimension $(p+1) \times 1$, its first component is $g(x)$ and the remaining $p$ components are the first derivatives of $g(x)$.

The point-wise local linear k-nn estimator of $\delta(x) = (g(x), \nabla g(x)')'$, for some $x \in R^p$, is given by

$$\hat{\delta}(x) = \left[\sum_{i=1}^{n} w_{R_n,x_i,x} \begin{pmatrix} 1, & (X_i - x)' \\ X_i - x, & (X_i - x)(X_i - x)' \end{pmatrix}\right]^{-1} \sum_{i=1}^{n} w_{R_n,x_i,x} \begin{pmatrix} 1 \\ X_i - x \end{pmatrix} Y_i, \tag{4.11}$$

where $W_{R_n,x_i,x} = R_n^{-p} W(\frac{X_i - x}{R_n})$.

The leave-one-out k-nn estimator for $\delta(X_i)$ which is given by

$$\hat{\delta}_{-i}(X_i) = \left[ \sum_{j \neq i}^{n} w_{R_i,ij} \begin{pmatrix} 1, & (X_j - X_i)' \\ X_j - X_i, & (X_j - X_i)(X_j - X_i)' \end{pmatrix} \right]^{-1} \sum_{j \neq i}^{n} w_{R_i,ij} \begin{pmatrix} 1 \\ X_j - X_i \end{pmatrix} Y_j, \tag{4.12}$$

where $w_{R_i,ij} = R_i^{-p} w(\frac{X_j - X_i}{R_i})$.

Define a $(p+1) \times 1$ vector $e_1$ whose first element is one with all remaining elements being zero. The leave-one-out k-nn estimator of $g(X_i)$ is given by $\hat{g}_{-i,L}(X_i) = e_1' \hat{\delta}_{-i}(X_i)$. We choose $k$ to minimize the following cross-validation objective function

$$CV_L(k) = n^{-1} \sum_{i=1}^{n} (Y_i - \hat{g}_{-i,L}(X_i))^2 M(X_i). \tag{4.13}$$

where $\hat{g}_{-i,L}(X_i) = e_1' \hat{\delta}_{-i}(X_i)$ and $M(.)$ is a weight function.

In D.2 we show that, uniformly in $k \in \Lambda$

$$CV_L(k) = \Phi_{1,L} \left( \frac{k}{n} \right)^{4/p} + \Phi_2 k^{-1} + o_p \left( \left( \frac{k}{n} \right)^{4/p} + k^{-1} \right),$$

where $\Phi_2$ is the same as defined in (4.8) and

$$\Phi_{1,L} = c_0^{-4/p} c_w^2 \int \left( \frac{1}{2} f(x) \; tr\left[ g_{ss}(x) \right] \right)^2 \left[ f(x) \right]^{-(p+4)/p} M(x) \, dx. \tag{4.14}$$

We observe that similar to the point-wise estimation result, the the local linear cross-validation objective function has a simple form for the leading bias term.

Let $k_{0,L}$ denote the value of $k$ that minimizes $\Phi_{1,L} \left( \frac{k}{n} \right)^{4/p} + \Phi_2 k^{-1}$, the leading term of $C_L(k)$, then it is easy to show that $k_{0,L} = a_L n^{4/(4+p)}$, where $a_L = [p\Phi_2/(4\Phi_{1,L})]^{p/(4+p)}$ is a positive constant. The next theorem shows that the cross validation selected $k$ is asymptotically equivalent to $k_{0,L}$.

**Theorem IV.2** *Under assumptions 1 to 3, and let $\hat{k}_L$ denote the local linear cross-*

*validation selected value of* $k$*, then*

$$\hat{k}_L = k_{0,L} + o\left(k_{0,L}\right) \text{ or equivalently } \hat{k}_L / k_{0,L} \xrightarrow{p} 1,$$

The proof of Theorem IV.2 is given in the D.2.

## D. Proofs

### 1. Proof of Theorem IV.1

Throughout section D, we write $A_n = B_n + (s.o.)$ to mean that the leading term of $A_n$ is $B_n$, $(s.o.)$ denotes terms having probability order smaller than that of $A_n$. When we write $A_n(X_i) = B_n(X_i) + (s.o.)$, the (s.o.) term is uniformly small for all values of $X_i \in \mathcal{S}$, where $\mathcal{S}$ is the support of the trimming function $M(.)$. Also, $A_n \sim B_n$ means that $A_n$ has the same probability order as that of $B_n$.

Let $S_r = \{v : ||v - x|| < r\}$ (a ball centered at $x$ with radius $r$), $G(r) = Prob[X_i \in S_r]$, $S_n = \{v : ||v - x|| < R_n\}$ and $P(S_n) = Prob[X_i \in S_n]$. Obviously $G(R_n) = P(S_n)$.

**Lemma IV.1** *Let* $\Phi(r) = 1/[r^\lambda G^\gamma(r)]$*,* $\lambda$ *and* $\gamma$ *are integers such that* $E[\Phi(R_n)]$ *exists, then*

$$E[\Phi(R_i)|X_i] = (c_0 f(X_i))^{\lambda/p} \left(\frac{k}{n}\right)^{-\lambda/p-\gamma} + (s.o.) \tag{4.15}$$

*where* $c_0 = \pi^{p/2}/\Gamma((p+2)/2)$ *is the volume of unit ball in* $R^p$*, and* $(s.o.)$ *denotes terms having probability order smaller than* $(k/n)^{-\lambda/p-\gamma}$*.*

Proof: Using equation (12) of Mack and Rosenblatt (1979), Liu and Lu (1997) have shown that (see lemma 1 of Liu and Lu) for $m = (\lambda + \eta)/p$, where $m$ is an integer and $\eta$ is a nonnegative integer less than or equal to $p - 1$ $(0 \leq \eta \leq p - 1)$,

$$E[\Phi(R_i)|X_i] = (c_0 f(X_i))^{\lambda/p} \frac{n!}{k!} \frac{(k-m-\gamma)!}{(n-m-\gamma)!} \left(\frac{(k-m-\gamma)}{(n-m-\gamma)}\right)^{\eta/p} + (s.o.). \tag{4.16}$$

Note that $\frac{n!}{k!}\frac{(k-m-\gamma)!}{(n-m-\gamma)!}\left(\frac{(k-m-\gamma)}{(n-m-\gamma)}\right)^{\eta/p} = \left(\frac{n}{k}\right)^{m+\gamma}\left(\frac{k}{n}\right)^{\eta/p} + (s.o.) = \left(\frac{k}{n}\right)^{\eta/p-m-\gamma} + (s.o.) = \left(\frac{k}{n}\right)^{-\lambda/p-\gamma} + (s.o.)$. Substituting this into (4.16) proves lemma IV.1.

**Lemma IV.2** *Let $A(x)$ be a measurable function of $x$. Then*

$$E[A(X_j)w(\frac{X_j - X_i}{R_i})|X_i, R_i] = \frac{k-1}{n}\frac{1}{G(R_i)}\int_{||x_j - X_i|| < R_i} f(x_j)A(x_j)w(\frac{x_j - X_i}{R_i})dx_j. \tag{4.17}$$

Proof: It follows directly from equation (22) of Mark and Rosenblatt (1979) and the fact that $w(\frac{x_j - X_i}{R_i}) = 0$ for $||x_j - X_i|| \geq R_i$.

**Proof of Theorem IV.1**

Using $Y_i = g(X_i) + u_i$, we have ($M_i = M(X_i)$)

$$\begin{aligned} CV(k) &= \frac{1}{n}\sum_{i=1}^{n}\left[g\left(X_i\right) - \hat{g}_{-i}\left(X_i\right)\right]^2 M_i \\ &\quad + \frac{2}{n}\sum_{i=1}^{n}\left[g\left(X_i\right) - \hat{g}_{-i}\left(X_i\right)\right] u_i M_i + \frac{1}{n}\sum_{i=1}^{n} u_i^2 M_i. \end{aligned} \tag{4.18}$$

The last term on the right of (4.18) is unrelated to $k$. Also, in lemma IV.6 below we show that the second term on the right of (4.18) has a probability order smaller than that of the first term on the right of (4.18). We denote this leading term by $CV_1(k)$ which is given by

$$\begin{aligned} CV_1(k) &= \frac{1}{n}\sum_{i=1}^{n}\left[g\left(X_i\right) - \hat{g}_{-i}\left(X_i\right)\right]^2 M(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[g\left(X_i\right) - \hat{g}_{-i}\left(X_i\right)\right]^2 \hat{f}_{-i}(X_i)^2 M(X_i)/\hat{f}_{-i}(X_i)^2 \end{aligned}$$

In lemma IV.4 below we further show that the leading term of $CV_1(k)$ is given by $CV_2(k)$, where $CV_2(k)$ is obtained from $CV_1(k)$ with $\hat{f}_{-i}(X_i)^{-2}$ being replaced by

$f(X_i)^{-2}$, i.e., $CV_1(k) = CV_2(k) + (s.o.)$, where

$$CV_2(k) = n^{-1} \sum_{i=1}^{n} \left[ g\left(X_i\right) - \hat{g}_{-i}\left(X_i\right) \right]^2 \hat{f}_{-i}(X_i)^2 M\left(X_i\right) / f(X_i)^2. \tag{4.19}$$

Since $\hat{g}_{-i}(X_i)\hat{f}_{-i}(X_i)$ contains one summation, $CV_2(k)$ contains three summations and therefore it can be written as a third order U-statistic, in lemma IV.6 we show that, using the U-statistic H-decomposition, the leading term of $CV_2(k)$ is $E[CV_2(k)]$. Define

$$\hat{m}_{1i} = \frac{1}{(n-1)R_i^p} \sum_{j \neq i}^{n} \left( g\left(X_j\right) - g\left(X_i\right) \right) w(\frac{X_j - X_i}{R_i}),$$

and

$$\hat{m}_{2i} = \frac{1}{(n-1)R_i^p} \sum_{j \neq i}^{n} u_j w(\frac{X_j - X_i}{R_i}).$$

Then $(\hat{g}_{-i}(X_i) - g(X_i))\hat{f}_{-i}(X_i) = \hat{m}_{1i} + \hat{m}_{2i}$, and from (4.19), we have

$$\begin{aligned} E[CV_2(k)] &= E[(\hat{m}_{1i} + \hat{m}_{2i})^2 f(X_i)^{-2} M(X_i)] \\ &= E[\hat{m}_{1i}^2 f(X_i)^{-2} M(X_i)] + E[\hat{m}_{2i}^2 f(X_i)^{-2} M(X_i)] \end{aligned} \tag{4.20}$$

because $E[\hat{m}_{1i}\hat{m}_{2i} f(X_i)^{-2} M(X_i)] = 0$.

We consider the two terms in (4.20) separately. By lemma IV.2 we know that

$$\begin{aligned} & E\left(\hat{m}_{i1}^2 | X_i = x, R_i = r\right) \\ = & \frac{1}{\left(n-1\right)^2 r^{2p}} E\left[ \left( \sum_{j \neq i}^{n} \left(g\left(X_j\right) - g\left(x\right)\right) w\left(\frac{X_j - x}{r}\right) \right)^2 | X_i = x, R_i = r \right] \\ = & \frac{(k-1)(k-2)}{\left(n-1\right)^2 r^{2p} G(r)^2} \left( \int f\left(x_j\right) \left(g\left(x_j\right) - g\left(x\right)\right) w\left(\frac{x_j - x}{r}\right) dx_j \right)^2 \\ & + \frac{(k-1)}{(n-1)^2 r^{2p} G(r)} \int f\left(x_j\right) \left(g\left(x_j\right) - g\left(x\right)\right)^2 w^2\left(\frac{x_j - x}{r}\right) dx_j \\ \equiv & A_{1n}(x,r) + A_{2n}(x,r), \end{aligned} \tag{4.21}$$

where the definition of $A_{jn}(x,r)$ should be apparent $(j=1,2)$. Note hat

$$\begin{aligned}
&\int f(x_j)\left(g(x_j)-g(x)\right)w\left(\frac{x_j-x}{r}\right)dx_j \\
=\;& r^p\int f(x+rv)\left(g(x+rv)-g(x)\right)w(v)\,dv \\
=\;& r^p\int\left(f(x)+rv'\nabla f(x)\right)\left(\nabla g(x)\,rv+\frac{1}{2}r^2v'\nabla^2g(x)\,v\right)w(v)\,dv+(s.o.) \\
=\;& r^{p+2}c_w\left(f(x)\,tr[\nabla^2g(x)]/2+\nabla f(x)'\nabla g(x)\right)+(s.o.), \qquad (4.22)
\end{aligned}$$

where $c_w=\int W(v)v_s^2dv$ is defined in assumption 2, and that

$$\begin{aligned}
&\int f(x_j)\left(g(x_j)-g(x)\right)^2w^2\left(\frac{x_j-x}{r}\right)dx_j \qquad (4.23)\\
=\;& r^p\int f(x+rv)\left(g(x+rv)-g(x)\right)^2w^2(v)\,dv \\
=\;& r^{p+2}\nu_w\int f(x)[\sum_{s=1}^{p}g_s(x)]^2+(s.o.), \qquad (4.24)
\end{aligned}$$

where $\nu_w=\int w^2(v)v_s^2dv$.

By (4.21) and (4.22) we obtain

$$\begin{aligned}
&E[A_{1n}(X_i,R_i)|X_i] \\
=\;& c_w^2\frac{k^2}{n^2}\left[f(x)tr[\nabla^2g(x)]/2+\nabla f(x)'\nabla g(x)\right]^2E[\frac{R_i^4}{G(R_i)^2}|X_i] \\
=\;& c_w^2\left(\frac{1}{c_0\,f(X_i)}\right)^{4/p-1}\left[f(X_i)tr[\nabla^2g(X_i)]/2+\nabla f(X_i)'\nabla g(X_i)\right]^2\left(\frac{k}{n}\right)^{4/p} \\
&+(s.o.), \qquad (4.25)
\end{aligned}$$

by lemma IV.1 since $E(\frac{1}{R_i^{-4}G(R_i)^2}|X_i)=\left[\frac{1}{c_0f(X_i)}\right]^{2/p}(k/n)^{4/p-2}+(s.o.)$.

By (4.21) and (4.23) we have

$$\begin{aligned} E[A_{2n}(X_i,R_i)|X_i] &= \frac{k}{n^2}c_w f(X_i)[\sum_{s=1}^{p} g_s(x)^2]E[\frac{1}{R_i^{p-2}G(R_i)}|X_i] \\ &= c_w f(X_i)[\sum_{s=1}^{p} g_s(x)^2]\left(\frac{1}{c_0 f(X_i)}\right)^{2/p-1}\frac{1}{k}\left(\frac{k}{n}\right)^{2/p} \\ &\quad +(s.o.), \end{aligned} \tag{4.26}$$

by lemma IV.1 because $E(\frac{1}{R_i^{p-2}G(R_i)}|X_i) = \left[\frac{1}{c_0 f(X_i)}\right]^{2/p-1}(k/n)^{2/p} + (s.o.)$.

Substituting (4.25) and (4.26) into (4.21) we obtain

$$\begin{aligned} & E\left(\hat{m}_{i1}^2|X_i\right) \\ = \; & E\left[E\left(\hat{m}_{i1}^2|X_i,R_i\right)|X_i\right] \\ = \; & \left[\frac{1}{cf\left(X_i\right)}\right]^{4/p}\left(c_w\sum\nolimits_{s=1}^{p}\left[\frac{1}{2}f\left(X_i\right)g_{ss}\left(X_i\right)+f_s\left(X_i\right)g_s\left(X_i\right)\right]\right)^2\left(\frac{k}{n}\right)^{4/p} \\ & +c_w f\left(X_i\right)\sum_{s=1}^{p}\left[g_s\left(X_i\right)\right]^2\left(\frac{1}{c\,f\left(X_i\right)}\right)^{2/p-1}\frac{1}{k}\left(\frac{k}{n}\right)^{2/p}+\left(s.o.\right). \end{aligned} \tag{4.27}$$

Similarly, by using lemma IV.2 we have

$$\begin{aligned} & E\left(\hat{m}_{i2}^2|X_i=x,R_i=r\right) \\ = \; & \frac{1}{\left(n-1\right)^2 r^{2p}}E\left(\sum_{j\neq i}^{n}u_j^2 w^2\left(\frac{X_j-X_i}{R_i}\right)|X_i=x,R_i=r\right) \\ = \; & \frac{k-1}{\left(n-1\right)^2 r^{2p}G(r)}\int\left(\sigma^2(x_j)w\left(\frac{X_j-x}{r}\right)\right)^2 dx_j \\ = \; & \frac{k-1}{\left(n-1\right)^2 r^{p}G(r)}\int\sigma^2\left(x+rv\right)w\left(v\right)^2 f\left(x+rv\right)r^p dv \\ = \; & \frac{k-1}{\left(n-1\right)^2 r^{p}G(r)}\sigma^2\left(x\right)f\left(x\right)\int w^2\left(v\right)dv+\left(s.o.\right) \\ = \; & \frac{d_w(k-1)}{\left(n-1\right)^2 r^{p}G(r)}\sigma^2\left(x\right)f\left(x\right)+\left(s.o.\right), \end{aligned} \tag{4.28}$$

where $d_w = \int w^2(v)dv$.

Therefore,

$$\begin{aligned} & E\left(\hat{m}_{i2}^2|X_i\right) = E\left[E\left(\hat{m}_{i2}^2|X_i,R_i\right)|X_i\right] \\ = & \frac{d_w(k-1)}{(n-1)^2}\sigma^2\left(X_i\right)f\left(X_i\right)E\left(\frac{1}{R_i^pG(R_i)}|X_i\right)+(s.o.) \\ = & d_w\sigma^2\left(X_i\right)f\left(X_i\right)\left[c_0f\left(X_i\right)\right]\frac{1}{k}+(s.o.) \\ = & c_0\,d_w\sigma^2\left(X_i\right)f^2\left(X_i\right)\frac{1}{k}+(s.o.) \end{aligned} \tag{4.29}$$

where we have used $E(\frac{1}{R_i^pG(R_i)}|X_i) = [c_0f\left(X_i\right)]\left(n/k\right)^2+(s.o.)$ by lemma IV.1.

Substituting (4.27) and (4.29) into (4.20) we obtain

$$\begin{aligned} & E[CV_2(k)] = E\left\{\left(\hat{m}_{i1}^2+\hat{m}_{i2}^2\right)f\left(X_i\right)^{-2}M\left(X_i\right)\right\} \\ = & \left\{c_0^{-4/p}c_w^2\int\left(\left[\frac{1}{2}f\left(x\right)tr[\nabla^2g\left(x\right)]+\sum_{s=1}^{p}f_s\left(x_i\right)g_s\left(x\right)\right]\right)^2\right. \\ & \left.[f\left(x\right)]^{-(p+4)/p}M\left(x\right)dx\right\}\left(\frac{k}{n}\right)^{4/p} \\ & +\left\{c_0d_w\int\sigma^2\left(x\right)M\left(x\right)f\left(x\right)dx\right\}\frac{1}{k}+(s.o.) \\ = & \Phi_1\left(\frac{k}{n}\right)^{4/p}+\Phi_2\frac{1}{k}+(s.o.), \end{aligned} \tag{4.30}$$

where $\Phi_1$ and $\Phi_2$ are two constants as defined in (4.7) and (4.8). Therefore, by (4.30), lemmas IV.4, IV.5 and IV.6 we know uniformly in $k\in\Lambda$ that

$$CV(k) = \Phi_1\left(\frac{k}{n}\right)^{4/p}+\Phi_2\frac{1}{k}+(s.o.). \tag{4.31}$$

Let $k_0$ denote the value of $k$ that minimizes $\Phi_1\left(\frac{k}{n}\right)^{4/p}+\Phi_2\frac{1}{k}$, then

$$k_0 = a_0n^{4/(4+p)}, \tag{4.32}$$

where $a_0 = [p\Phi_2/(4\Phi_1)]^{p/(4+p)}$. From (4.31) we know that $\hat{k} = k_0+o\left(k_0\right)$, or equivalently, $\hat{k}/k_0\xrightarrow{p}1$ This completes the proof of Theorem IV.1.

$(0 < \delta < 1)$ such that (for $j = 1, 2$)

Below we present and prove lemma IV.3 to lemma IV.6 that are used in proving Theorems IV.1 and IV.2.

**Lemma IV.3** *Denotes by* $R_n = R_n(x)$ *and define*
$\hat{m}_{1,-i}(x) = \frac{1}{(n-1)R_n^p} \sum_{j \neq i}^{n} (g(X_j) - g(x)) w\left(\frac{X_j - x}{R_n}\right)$,
$\hat{m}_{2,-i}(x) = \frac{1}{(n-1)R_n^p} \sum_{j \neq i}^{n} u_j w\left(\frac{X_j - x}{R_n}\right)$.
*Define* $\Delta_{m1,-i} = \hat{m}_{1,-i}(x) - E(\hat{m}_{1,-i}(x))$, *and* $\Delta_{f,-i} = \hat{f}_{-i}(x) - E\left(\hat{f}_{-i}(x)\right)$.
*Then for* $j = 1, 2$, *we have*

$$\sup_{(x,k)} \left| \frac{1}{n} \sum_{i=1}^{n} \hat{m}_{j,-i}^2 / \hat{f}(x)^2 M(x) - \frac{1}{n} \sum_{i=1}^{n} \hat{m}_{j,-i}^2 / f(x)^2 M(x) \right| = o_p\left( \frac{1}{k} + \left(\frac{k}{n}\right)^{4/p} \right),$$

*where the supreme is over* $(x, k) \in \mathcal{S} \times \Lambda$ *(*$\mathcal{S}$ *is the support of* $M(.)$*).*

Proof: The proof follows the same arguments for $j = 1$ and $j = 2$. Therefore, we only

prove the case of $j = 1$. By Rosenthal's inequality we have

$$\begin{aligned}
& E\left\{\left|\Delta_{m1,-i}\right|^{2t}\right\} \\
= \; & E\left\{\left|\frac{1}{n}\sum_{j=1}^{n}\left[\frac{1}{R_n^p}\left(g\left(X_j\right)-g\left(x\right)\right)w\left(\frac{X_j-x}{R_n}\right)-\mu_{m1}\right]\right|^{2t}\right\} \\
\leq \; & n^{-2t}C_t\left\{n^t\left(E\left[\frac{1}{R_n^p}\left(g\left(X_j\right)-g\left(x\right)\right)w\left(\frac{X_j-x}{R_n}\right)-\mu_{m1}\right]^2\right)^t\right. \\
& \left.+n\,E\left[\left|R_n^{-p}\left(g\left(X_j\right)-g\left(x\right)\right)w\left(\frac{X_j-x}{R_n}\right)\right|^{2t}\right]\right\} \\
\sim \; & n^{-2t}C_t\left\{n^t\left(E\left[\frac{k}{n}\frac{1}{R_n^p}\left(g\left(X_j\right)-g\left(x\right)\right)w\left(\frac{X_j-x}{R_n}\right)\right]^2\right)^t\right\} \\
= \; & n^{-2t}C_t\left\{k^t\left(\int\left[\frac{f(x_j)}{R_n^pG(R_n)}\left(g\left(x_j\right)-g\left(x\right)\right)w\left(\frac{x_j-x}{R_n}\right)\right]^2dx_j\right)\right\} \\
\sim \; & n^{-2t}C_t\left\{k^t\int\left[\left(\frac{f(x+R_nv)}{G(R_n)}\left(g\left(x+R_nv\right)-g\left(x\right)\right)w\left(v\right)\right]^2dv\right)\right\} \\
\sim \; & n^{-2t}C_t\left\{k^t\left(\frac{k}{n}\right)^{\left(\frac{2}{p}-2\right)t}\right\} \\
= \; & O\left(\frac{1}{k^t}\left(\frac{k}{n}\right)^{\frac{2t}{p}}\right).
\end{aligned}$$

By Markov's inequality and note that $n^\epsilon < k < n^{1-\epsilon}$, then we have

$$\begin{aligned}
& P\left\{\left|\Delta_{m1}\right| > n^{-\delta}k^{-1/3}\right\} \\
\leq \; & Cn^{2t\delta}k^{2t/3}\frac{1}{k^t}\left(\frac{k}{n}\right)^{\frac{2t}{p}} = Cn^{t\left(2\delta-\frac{2}{p}\right)}k^{t\left(\frac{2}{3}-1+\frac{2}{p}\right)} \\
\leq \; & C\left[n^{2\delta-\frac{2}{p}+(1-\epsilon)\left(-\frac{1}{3}+\frac{2}{p}\right)}\right]^t = C\left[n^{-\frac{1}{3}[(1-\epsilon)-6\delta]-\epsilon\frac{2}{p}}\right]^t \\
\leq \; & const.\,n^{-C} = O(n^{-C}), \qquad (4.33)
\end{aligned}$$

for some $0 < \delta < (1-\epsilon)/6$ and for all $\xi = (x,k) \in \mathcal{S} \times \Lambda$.

Since $C$ in (4.33) can be arbitrarily large, then the same result holds uniformly in

any set of size no larger than a polynomial in $n$ of values of $(x,k)$. Furthermore, the the weight function $M(.)$ is compactly supported and Hölder continuous. Therefore, taking a polynomially fine mesh of vectors $\xi \in \mathcal{S} \times \Lambda$, we deduce that (4.33) continue to hold when taking supreme over $(x,k) \in \mathcal{S} \times \Lambda$. Thus, we have that

$$P\left(sup_{(x,k)}\ k^{1/3} \left|\Delta_{m1,-i}\right| > n^{-\delta}\right) = O(n^{-c}), \tag{4.34}$$

which implies that $\Delta_{m1}(x) = o_p(k^{-1/3})$ uniformly in $(x,k) \in \mathcal{S} \times \Lambda$. By exactly the same arguments, one can show that

$$\sup_{(x,k)} \left|\Delta_{f,-i}\right| = o_p\left(k^{-1/3}\right). \tag{4.35}$$

By the fact that the derivative functions $g_{ss}(.)$, $g_s(.)$ and $f_s(.)$ are all continuous and bounded, and the simple change of variable and Taylor expansion argument leads to

$$\sup_{(x,k)} \left[E\left(\hat{m}_{i1}\left(x\right)\right)\right] = O\left(\left(\frac{k}{n}\right)^{2/p}\right) \tag{4.36}$$

Hence,

$$\begin{aligned} \sup_{(x,k)} \left|\hat{m}_{1,-i}\left(x\right)\right| &\leq \sup_{x} \left|\hat{m}_{1,-i}\left(x\right) - E\left(\hat{m}_{1,-i}\left(x\right)\right)\right| + \sup_{(x,k)} \left|E\left(\hat{m}_{1,-i}\left(x\right)\right)\right| \\ &= o_p\left(k^{-1/3}\right) + O_p\left(\frac{k}{n})^{2/p}\right). \end{aligned} \tag{4.37}$$

Similar arguments lead to

$$\sup_{(x,k)} \left|\hat{f}(x) - f(x)\right| = o_p\left(k^{-1/3}\right) + O_p\left(\left(\frac{k}{n}\right)^{2/p}\right). \tag{4.38}$$

From (4.37), (4.38) and the assumption that $inf_{x\in\mathcal{S}} f(x) \geq \delta > 0$, we reduce that

$$\begin{aligned} & \sup_{(x,k)} \left| \frac{1}{n} \sum_{i=1}^{n} \hat{m}_{1,-i}^2 / \hat{f}\left(x\right)^2 M\left(x\right) - \frac{1}{n} \sum_{i=1}^{n} \hat{m}_{1,-i}^2 / f\left(x\right)^2 M\left(x\right) \right| \\ \leq \; & C \sup_{(x,k)} \left| \hat{m}_{1,-i}^2 \right| \sup_{(x,k)} \left| \hat{f}\left(x\right) - f\left(x\right) \right| = o_p\left( \frac{1}{k} + \left( \frac{k}{n} \right)^{4/p} \right). \end{aligned}$$

**Lemma IV.4** *$CV_1(k) = CV_2(k) + o\left((k/n)^{4/p} + k^{-1}\right)$, where $o(.)$ is uniformly in $k$.*

Proof: It follows directly from lemma IV.3.

**Lemma IV.5** *$CV_2(k) = E[CV_2(k)] + o_p\left((k/n)^{4/p} + k^{-1}\right)$, where $o(.)$ is uniformly in $k \in \Lambda$.*

Proof: $CV_2(k)$ can be written as a third order U-statistic:

$CV_2(k) = \frac{6}{n(n-1)(n-2)} \sum\sum\sum_{l>j>i} H_{n,ijl}$, where $H_{n,ijl} = H_n(z_i, z_j, z_l)$, is the symmetrized version of

$$\left[g\left(X_i\right) - g\left(X_j\right)\right] R_{W_i,ij} \left[g\left(X_i\right) - g\left(X_l\right)\right] R_{W_i,il} M\left(X_i\right) / f(X_i)^2,\ z_i = (X_i, R_i).$$

Denotes by $\theta = E[H_{n,ijl}]$, $H_{n,ij} = E[H_{n,ijl}|z_i, z_j]$ and $H_{ni} = E[H_{n,ij}|z_i]$. Then using the U-statistic H-decomposition we have

$$\begin{aligned} CV_2(k) &= \theta + \frac{3}{n} \sum_i \left[H_{ni} - \theta\right] + \frac{6}{n(n-1)} \sum\sum_{j>i} \left[H_{n,ij} - H_{ni} - H_{nj} + \theta\right] \\ &+ \frac{6}{n(n-1)(n-2)} \sum\sum\sum_{l>j>i} [H_{n,ijl} - H_{n,ij} \\ &\quad - H_{n,jl} - H_{n,il} + H_{ni} + H_{nj} + H_{nl} - \theta] \\ &\equiv \theta + A_{1n} + A_{2n} + A_{3n}, \end{aligned} \tag{4.39}$$

where $A_{jn}$ contains $j$ summations ($j = 1, 2, 3$). and their specific definitions should apparent from (4.39).

Note that $\theta = E[CV_2(k)]$, so we only need to show $A_{jn} = o_p((k/n)^{4/p} + k^{-1})$ uniformly in $k \in \Lambda$ for $j = 1, 2, 3$. Note that $E[H_n^2(z_i)] = O((k/n)^{4/p})$. By Rosenthal's

inequality,

$$E|A_{1n}|^{2t} \leq n^{-2t} C_t \left[ n^t ((k/n)^{4/p})^{2t} + (s.o.) \right] = O(n^{-t}(k/n)^{8t/p}).$$

Then by Markov's inequality

$$P\left( |A_{1n}| > n^{-\delta}(k/n)^{4/p} \right) \leq C_t n^{-(1-2\delta)t} = O(n^{-C})$$

uniformly in $k$. Since $C$ is arbitrarily large, this implies that

$$P\left( \sup_{k \in \Lambda} (k/n)^{-4/p} |A_{1n}| > n^{-\delta} \right) = O(n^{-C})$$

which implies that $A_{1n} = o_p((k/n)^{4/p})$ uniformly in $k \in \Lambda$. Using similar arguments one can show that $A_{jn} = o_p((k/n)^{4/p})$ uniformly in $k \in \Lambda$ for $j = 2, 3$.

**Lemma IV.6** *$CV(k) = CV_1(k) + o_p((k/n)^{4/p} + k^{-1})$ + a term not related to $k$, where the $o_p(.)$ term is uniformly in $k \in \Lambda$.*

Let $A_{4n} = n^{-1} \sum_i u_i (g_i - \hat{g}_{-i}) M_i$. Then by equation (4.18) it suffices to show that $A_{4n} = o_p\left( (k/n)^{4/p} + k^{-1} \right)$ uniformly in $k$. Similar to the proof of lemma IV.4 one can show that

$A_{4n} = n^{-1} \sum_i u_i (g_i - \hat{g}_{-i}) \hat{f}_{-i} M_i / \hat{f}_{-i} = n^{-1} \sum_i u_i (g_i - \hat{g}_{-i}) \hat{f}_{-i} M_i / f_i + (s.o.) \equiv A_{4n,1} + (s.o.)$.

That is, the leading term of $A_{4n}$ is $A_{4n,1}$ which is obtained by replacing $\hat{f}_i^{-1}$ in $A_{4n}$ by $f_i^{-1}$. Then using $Y_j = g_j + u_j$ we can write $A_{4n,1}$ as $A_{4n,1} = B_{1n} + B_{2n}$, where $B_{1n} = [n(n-1)]^{-1} \sum_i \sum_{j \neq i} u_i (g_i - g_j) w_{R_i,ij} M_i / f_{-i}$ and $B_{2n} = [n(n -$

$1)]^{-1}\sum_i\sum_{j\neq i}u_iu_jw_{R_i,ij}M_i/f_{-i}$. We consider $B_{1n}$ and $B_{2n}$ separately below.

$$\begin{aligned}
E[B_{1n}^2] &= \frac{1}{n^2(n-1)^2}\sum_i\sum_{j\neq i}\sum_{l\neq i}E\left[u_i^2(g_i-g_j)w_{R_i,ij}(g_i-g_l)w_{R_i,il}M_i^2/f_i^2\right]\\
&= \frac{1}{n^2(n-1)^2}\sum_i\sum_{j\neq i}E\left[\sigma^2(X_i)(g_i-g_j)^2w_{R_i,ij}^2M_i^2/f_i^2\right]\\
&\quad+\frac{1}{n^2(n-1)^2}\sum_i\sum_{j\neq i}\sum_{l\neq i,l\neq i}E\left[\sigma^2(X_i)(g_i-g_j)w_{R_i,ij}(g_i-g_l)w_{R_i,il}M_i^2/f_i^2\right]\\
&\equiv B_{1n,1}+B_{1n,2}.
\end{aligned}$$

Using $E\left[(g_i-g_j)^2w_{R_i,ij}^2|X_i,R_i\right]=O(R_i^2/R_i^p)$, it is easy to easy to see that $B_{1n,1}=n^{-2}O(E[R_i^2/R_i^p])=n^{-2}O((k/n)^{2/p-1})=O((nk)^{-1}(k/n)^{2/p})$.

Next, conditional on $(X_i,R_i)$ $E[(g_i-g_j)w_{R_i,ij}|X_i,R_i]=O(R_i^2)$. Hence, we have $B_{1n,2}=n^{-1}O(E(R_i^4))=O(n^{-1}(k/n)^{4/p})$.

Thus, we have shown that $E[B_{1n}^2]=B_{1n,1}+B_{1n,2}=O((k/n)^{4/p}n^{-1}+(nk)^{-1}(k/n)^{2/p})$ and this implies that $B_{1n}=O_p((k/n)^{2/p}n^{-1/2}+(nk)^{-1/2}(k/n)^{2/p})$.

Note that $B_{2n}$ can be written as

$$B_{2n}=\frac{1}{n(n-1)}\sum_i\sum_{j>i}u_iu_j\left[W_{R_i,ij}+W_{R_j,ji}\right]$$

We have

$$\begin{aligned}
E[B_{2n}^2] &= \frac{1}{n^2(n-1)^2}\sum_i\sum_{j>i}E\left\{u_i^2u_j^2\left[W_{R_i,ij}+W_{R_j,ji}\right]^2\right\}\\
&= \frac{1}{n^2(n-1)^2}\sum_i\sum_{j>i}E\left\{\sigma^2(X_i)\sigma^2(X_j)\left[W_{R_i,ij}+W_{R_j,ji}\right]^2\right\}\\
&\leq \frac{4}{n^2(n-1)^2}\sum_i\sum_{j>i}E\left[\sigma^2(X_i)\sigma^2(X_j)W_{R_i,ij}^2\right]\\
&= n^{-2}O(E[R_i^{-p}])=O(n^{-2}(k/n)^{-1})=O((nk)^{-1}).
\end{aligned}$$

Hence, $B_{2n}=O_p((nk)^{-1/2})$.

Summarizing the above we have shown that $B_{1n}+B_{2n}=O_p((k/n)^{2/p}n^{-1/2}+$

$(nk)^{-1/2}) = o_p((k/n)^{4/p} + k^{-1})$. It can be shown that the result holds uniformly in $k \in \Lambda$. Hence, $A_{4n,1} = o_p((k/n)^{4/p} + k^{-1})$ uniformly in $k \in \Lambda$ which in turn gives that $CV(k) - CV_1(k) = A_{4n} +$ a term unrelated to k $= o_p((k/n)^{4/p} + k^{-1}) +$ a term unrelated to k. This completes the proof of lemma IV.6.

## 2. Proof of Theorem IV.2

The cross-validation function is

$$CV_L(k) = n^{-1} \sum_{i=1}^{n} (Y_i - \hat{g}_{-i,L}(X_i))^2 M(X_i)$$

where $\hat{g}_{-i,L}(X_i)$ is the leave-one-out local linear k-nn estimator of $g(X_i)$.

Insert the identity matrix $I_{p+1} = \mathcal{G}_n^{-1}\mathcal{G}_n$ in the middle of (4.12), where $\mathcal{G}_n = \begin{pmatrix} R_i^2 & 0 \\ 0 & I_p \end{pmatrix}$, we have

$$\begin{aligned}
\hat{\delta}_{-i}(X_i) &= \left[ \sum_{j \neq i} w_{R_i,ij} \mathcal{G}_n \begin{pmatrix} 1 \\ X_j - X_i \end{pmatrix} (1, (X_j - X_i)') \right]^{-1} \\
&\quad \times \sum_{j \neq i} w_{R_i,ij} \mathcal{G}_n \begin{pmatrix} 1 \\ X_j - X_i \end{pmatrix} Y_j \\
&= \left[ \sum_{j \neq i} w_{R_i,ij} \begin{pmatrix} R_i^2 \\ X_j - X_i \end{pmatrix} (1, (X_j - X_i)') \right]^{-1} \\
&\quad \times \sum_{j \neq i} w_{R_i,ij} \begin{pmatrix} R_i^2 \\ X_j - X_i \end{pmatrix} Y_j. \qquad (4.40)
\end{aligned}$$

The advantage of using (4.40) in the proof is that, conditional on $X_i$, $[\frac{1}{nR_i^{p+2}} \sum_{j \neq i} W(\frac{X_j - X_i}{R_i}) \begin{pmatrix} R_i^2 \\ X_j - X_i \end{pmatrix} (1, (X_j - X_i)')]$ converge to a non-singular ma-

trix, which simplifies the proof significantly.

By Taylor series expansion: $g(X_j) = g(X_i) + (X_j - X_i)'\nabla g(X_i) + T_{ij}$, where $T_{ij} = g(X_j) - g(X_i) - (X_j - X_i)'\nabla g(X_i)$ is the remainder term in the Taylor expansion. Then (4.40) leads to

$$\begin{aligned}
\hat{\delta}_{-i}(X_i) &= \left[\frac{1}{nR_i^2}\sum_{j\neq i} w_{R_i,ji}\begin{pmatrix} R_i^2 \\ X_j - X_i \end{pmatrix}(1, (X_j - X_i)')\right]^{-1} \\
&\quad \times \left\{\frac{1}{nR_i^2}\sum_{j\neq i} w_{R_i,ji}\begin{pmatrix} R_i^2 \\ X_j - X_i \end{pmatrix}[g(X_j) + u_j]\right\} \\
&= \delta(X_i) + [\frac{1}{nR_i^2}\sum_{j\neq i} w_{R_i,ji}\begin{pmatrix} R_i^2, & R_i^2(X_j - X_i)' \\ X_j - X_i, & (X_j - X_i)(X_j - X_i)' \end{pmatrix}]^{-1} \\
&\quad \times \{\frac{1}{nR_i^2}\sum_{j\neq i} w_{R_i,ji}\begin{pmatrix} R_i^2 \\ X_j - X_i \end{pmatrix}[T_{ij} + u_j]\} \\
&\equiv \delta(X_i) + A_{2i}^{-1}A_{1i}, \qquad (4.41)
\end{aligned}$$

where

$$A_{1i} = \frac{1}{nR_i^2}\sum_{j\neq i} w_{R_i,ji}\begin{pmatrix} R_i^2 \\ X_j - X_i \end{pmatrix}[T_{ij} + u_j],$$

and

$$A_{2i} = \begin{pmatrix} \hat{f}_i & B_{1i}' \\ R_i^{-2}B_{1i} & R_i^{-2}B_{2i} \end{pmatrix}$$

where

$$\hat{f}_i = n^{-1}\sum_{j\neq i} w_{R_i,ji},$$

$$B_{1i} = n^{-1}\sum_{j\neq i} w_{R_i,ji}(X_j - X_i),$$

$$B_{2i} = n^{-1}\sum_{j\neq i} w_{R_i,ji}(X_j - X_i)(X_j - X_i)'.$$

Using partitioned inverse, we obtain

$$e_1' A_{2i}^{-1} = (\hat{f}_i^{-1} + C_{1i}, -C_{2i}),$$

where

$$C_{1i} = \hat{f}_i^{-2} B_{1i}' [R_i^{-2}(B_{2i} - B_{1i}B_{1i}'\hat{f}_i^{-1})]^{-1} B_{1i},$$

and

$$C_{2i} = \hat{f}_i^{-1} B_{1i}' [R_i^{-2}(B_{2i} - B_{1i}B_{1i}'\hat{f}_i^{-1})]^{-1}.$$

Therefore,

$$\begin{aligned}
\hat{g}_{-i}(X_i) &= e_1'\hat{\delta}_{-i}(X_i) = g(X_i) + e_1' A_{2i}^{-1} A_{1i} = g(X_i) + (\hat{f}_i^{-1} + C_{1i}, -C_{2i}) A_{1i} \\
&= g(X_i) + n^{-1} \sum_{j \neq i} w_{R_i,ij}[T_{ij} + u_j]/\hat{f}_i \\
&\quad + n^{-1} \sum_{j \neq i} w_{R_i,ij}[T_{ij} + u_j][C_{1i} - R_i^{-2} C_{2i}(X_j - X_i)] \\
&\equiv \tilde{g}(X_i) + Q_i, \qquad (4.42)
\end{aligned}$$

where $\tilde{g}(X_i) = g(X_i) + n^{-1} \sum_{j \neq i} w_{R_i,ij}[T_{ij} + u_j]/\hat{f}_i$ and $Q_i = n^{-1} \sum_{j \neq i} w_{R_i,ij}[T_{ij} + u_j][C_{1i} - R_i^{-2} C_{2i}(X_j - X_i)]$.

Substituting (4.42) into (4.40) (and omitting $M(X_i)$ for notational simplicity), we have

$$CV_L(K) = \frac{1}{n} \sum_i (Y_i - \tilde{g}_i)^2 + \frac{2}{n} \sum_i (Y_i - \tilde{g}_i) Q_i + \frac{1}{n} \sum_i Q_i^2. \qquad (4.43)$$

In lemma IV.7 we show that the leading term of $CV_L(k)$ is the first term on the right side of (4.43), i.e., $CV_L(k) = CV_{L,1}(k) + (s.o.)$, where

$$\begin{aligned}
CV_{L,1}(k) &= n^{-1} \sum_i (Y_i - \tilde{g}_i)^2 = n^{-1} \sum_i (g_i + u_i - \tilde{g}_i)^2 \\
&= n^{-1} \sum_i (g_i - \tilde{g}_i)^2 - 2n^{-1} \sum_i u_i (g_i - \tilde{g}_i) + n^{-1} \sum_i u_i^2. \qquad (4.44)
\end{aligned}$$

The third term above does not depend on $k$, and by essentially the same arguments

as in the proof of Theorem IV.1 one can show that the second term on the right of (4.44) has an order smaller than the first term. Thus, the leading term of $CV_{L,1}(k)$ (hence $CV_L(k)$) is ($g_i = g(X_i)$ and $\tilde{g}_i = \tilde{g}(X_i)$)

$$CV_{L,0}(k) = n^{-1} \sum_i (g_i - \tilde{g}_i)^2. \tag{4.45}$$

By comparing (4.45) with the leading term for the local constant term case, we observe that the only difference is that $(g_i - \hat{g}_{-i}) = [(n-1)R_i^p]^{-1} \sum_{j \neq i} (g_i - g_j) w(\frac{X_j - X_i}{R_i}) / \hat{f}_{-i}$ in the local constant estimator case is now replaced by $g_i - \tilde{g}_i = [(n-1)R_i^p]^{-1} \sum_{j \neq i} (g_i - g_j - (X_j - X_i)' \nabla g(X_i)) w(\frac{X_j - X_i}{R_i}) / \hat{f}_{-i}$. This change will only affect the bias calculation in the local linear estimator. The leading bias term now only depends on $\nabla^2 g(.)$, not depends on $g_s(.) f_s(.)$. Therefore, by *exactly* the same derivations as in the proof of Theorem 2.1, one can show that the leading term of $CV_{L,0}(k)$ is $E[CV_{L,0}(k)]$ and that

$$E[CV_{L,0}(k)] = \Phi_{1,L} \left( \frac{k}{n} \right)^{4/p} + \Phi_2 \frac{1}{k} + o_p \left( k^{-1} + (k/n)^{4/p} \right). \tag{4.46}$$

Note that the only difference between $\Phi_{1,L}$ and $\Phi_1$ defined in Theorem IV.1 is that there is no $g_s(.) f_s(.)$ term in $\Phi_{1,L}$. This is an expected result as the leading bias term in the local linear estimation does not involve first partial derivative functions.

**Lemma IV.7** *$CV_L(k) = CV_{L,1}(k) + o_p((k/n)^{4/p} + k^{-1})$ uniformly in $k \in \Lambda \in \Lambda$.*

Proof: The proof is similar to the arguments we used in proving Theorem IV.1. A detailed proof is quite tedious, here we only provide a sketchy proof.

Note that $CV_{L,1}(k) = n^{-1} \sum_i [Y_i - \tilde{g}_i]^2$ (we omit the weighting function $M_i$) and

$$CV_L(k) = n^{-1} \sum_i [Y_i - \tilde{g}_i]^2 - 2n^{-1} \sum_i (Y_i - \tilde{g}_i) Q_i + n^{-1} \sum_i Q_i^2. \tag{4.47}$$

We need to show that both $n^{-1} \sum_i (Y_i - \tilde{g}_i) Q_i$ and $n^{-1} \sum_i Q_i^2$ are $o_p((k/n)^{4/p} + k^{-1})$.

It can be shown that uniformly in $i$, $B_{1i} = O_p(E(R_i^2)) = O_p((k/n)^{2/p})$ and $B_{2i} = O_p(E(R_i^2)) = O_p((k/n)^{2/p})$. From this one can further show that uniformly in $i$, $C_{1i} = O_p((k/n)^{2/p})$ and $C_{2i} = O_p((k/n)^{2/p})$. Now comparing $\tilde{g}(X_i) - g(X_i) = n^{-1}\sum_{j\neq i} w_{R_i,ij}[T_{ij} + u_j]/\hat{f}_i$ and $Q_i = n^{-1}\sum_{j\neq i} w_{R_i,ij}[T_{ij} + u_j][C_{1i}R_i^{-2}C_{2i}R_i\frac{X_j - X_i}{R_i}] = n^{-1}\sum_{j\neq i} w_{R_i,ij}[T_{ij}+u_j]C_{1i} - R_i n^{-1}\sum_{j\neq i}\bar{w}_{R_i,ij}[T_{ij}+u_j]R_i^{-2}C_{2i}$, where $\bar{w}_{R_i,ij} = (\frac{X_j - X_i}{R_i})$ $\times w_{R_i,ij}$ is a new weight function. Now if one replaces $C_{1i}$ and $C_{2i}$ by $R_i^2$, it should be obvious that $Q_i$ has an smaller order than $\tilde{g}(X_i) - g(X_i)$. This together with the fact that

$$n^{-1}\sum_i [\tilde{g}_i - g_i]^2 = O_p((k/n)^{4/p} + k^{-1}) \tag{4.48}$$

leads to

$$\frac{1}{n}\sum_i Q_i^2 = o_p((k/n)^{4/p} + k^{-1}). \tag{4.49}$$

(4.48), (4.49) and Cauchy inequality imply that

$$\frac{1}{n}\sum_i Q_i(\tilde{g}_i - g_i) = o_p((k/n)^{4/p} + k^{-1}). \tag{4.50}$$

Finally, it is not hard to show that

$$\frac{1}{n}\sum_i Q_i(Y_i - \tilde{g}_i) = \frac{1}{n}\sum_i Q_i(g_i - \tilde{g}_i) + (s.o.) = o_p((k/n)^{4/p} + k^{-1}), \tag{4.51}$$

where the last equality follows from (4.50).

Lemma IV.7 now follows from (4.47), (4.49) and (4.51).

# CHAPTER V

# UNIFORM CONVERGENCE RATE OF KERNEL ESTIMATION OF REGRESSION FUNCTIONS WITH MIXED CATEGORICAL AND CONTINUOUS DATA

## A. Introduction

It is well known that the traditional 'frequency-based' approach can be used to consistently estimate a nonparametric regression function with categorical variables. However, the use of this conventional frequency-based approach to handle the discrete variables is known to be unsatisfactory. Because many economic data contain a mixture of discrete and continuous variables, and when the sample size is not sufficiently large compared with the number of discrete cells, the frequency-based method cannot lead to accurate nonparametric estimation results. In a seminal paper Aitchison and Aitken (1976) propose an novel idea of smoothing the discrete variables. Recently, Hall, Racine and Li (2004), Li and Racine (2003), and Racine and Li (2004) have extended the idea of Aitchison and Aitken (1976) to the case with mixed discrete and continuous variables. They suggest to smooth both the discrete and continuous variables, and use the data-driven cross-validation method to select the smoothing parameters. Hall, Racine and Li (2004) have shown that the cross-validation method has the amazing ability of (asymptotically) automatically removing irrelevant variables. This is important for nonparametric estimation, since the 'curse of dimensionality' is the main obstacle of nonparametric estimation method. Li and Racine (2005) have shown that 'irrelevant variables' occurs surprisingly often in practice, and they show that, for a variety of economic data such as U.S. patent application data, U.S. crop yield insurance premium data, U.S. labor force participation data, U.S. inflation data,

and U.S. marketing data, by smoothing both the discrete and continuous variables often lead to superior out-of-sample predictions compared with some commonly used parametric approaches.

The aim of this chapter is to establish uniform convergence rates of nonparametric regression or density estimators in the presence of mixed discrete and continuous variables, we allow for deterministic or data-driven selected smoothing parameters. This is particularly important since data-driven method seems to be the only practical method to select the smoothing parameters in the mixed *categorical* and continuous variable case.

## B. Uniform Convergence Rate of Kernel Regression Estimator

We consider the following nonparametric regression model.

$$Y_i = g(x_i) + u_i, \qquad (i = 1, ..., n) \tag{5.1}$$

where $x_i = (x_i^c, x_i^d)$, $x^d$ is a discrete variable of dimension $r$, and takes values on $\mathcal{S}^d = \prod_{s=1}^{r}\{0, 1, ..., c_s - 1\}$, $x^c \in \mathcal{R}^q$ is a continuous variable. We use $x_{is}^c$ and $x_{is}^d$ to denote the $s$th component of $x_i^c$ and $x_i^d$, respectively. Denotes by $w(\frac{x_s - x_{is}^c}{h_s})$ the univariate kernel function for $x_s^c$, $h_s$ is the smoothing parameter. The product kernel for $x^c$ is given by

$$W_{h,x,x_i} = \prod_{s=1}^{q} \frac{1}{h_s} w(\frac{x_{is}^c - x_s^c}{h_s}). \tag{5.2}$$

We assume that some of the discrete variables do not have a natural ordering, and the remaining ones have a natural ordering (e.g. preference orderings: like, indifference, dislike). Let $\bar{x}_i^d$ denote a $r_1 \times 1$ vector regressors that do not have a natural ordering and let $\tilde{x}_i^d$ denote the remaining $r_2 = r - r_1$ discrete regressors that have a natural ordering.

For an unordered variable, we use the following kernel function

$$\bar{l}(\bar{x}_{is}^d, \bar{x}_{js}^d, \lambda_s) = \begin{cases} 1, & \text{if } \bar{x}_{is}^d = \bar{x}_{js}^d, \\ \lambda_s, & \text{otherwise,} \end{cases} \tag{5.3}$$

where $\lambda_s \in [0, 1]$ is the smoothing parameter.

For an ordered variable, the kernel function is defined by:

$$\tilde{l}(\tilde{x}_{is}^d, \tilde{x}_{js}^d, \lambda_s) = \begin{cases} 1, & \text{if } \tilde{x}_{is}^d = \tilde{x}_{js}^d, \\ \lambda_s^{|\tilde{x}_{is}^d - \tilde{x}_{js}^d|}, & \text{if } \tilde{x}_{is}^d \neq \tilde{x}_{js}^d. \end{cases} \tag{5.4}$$

In both (5.3) and (5.4) when $\lambda_s = 0$ $l(.,.,0)$ becomes an indicator function, and when $\lambda_s = 1$, $l(.,.,1) \equiv 1$ is a uniform weight function.

Let $\mathbf{1}(A)$ denote an indicator function that assumes the value 1 if $A$ occurs and 0 otherwise. Combining (5.3) and (5.4), we obtain the product kernel function given by

$$L_{\lambda,x,x_i} = \left[\prod_{s=1}^{r_1} \lambda_s^{\mathbf{1}(\bar{x}_s^d \neq \bar{x}_{is}^d)}\right] \left[\prod_{s=r_1+1}^{r} \lambda_s^{|\tilde{x}_s^d - \tilde{x}_{is}^d|}\right]. \tag{5.5}$$

The final product kernel for the mixed discrete and continuous variables is given by

$$K_{\gamma,x,x_i} = W_{h,x,x_i} L_{\lambda,x,x_i}$$

where $\gamma = (h, \lambda) = (h_1, ..., h_q, \lambda_1, ..., \lambda_r)$. We estimate $g(x)$ by

$$\hat{g}(x) = \frac{1}{n} \sum_{i=1}^{n} y_i K_{\gamma,x,x_i} / \hat{f}(x) \tag{5.6}$$

where $\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_{\gamma,x,x_i}$ is the kernel estimator of $f(x)$, the density function of $X$.

From (5.6) it is easy to see that if $\lambda_s = 1$ (the upper extreme value) for some $1 \leq s \leq r$, then the $x_s^d$ variable is completely smoothed out. Because in this case, $l(x_s^d, x_{is}^d, 1) \equiv 1$ for all values of $x_s^d$ and $x_{is}^d$, so that $x_s^d$ does not affect the nonparametric

estimator $\hat{g}(x)$.

The purpose of this chapter is to derive the uniform convergence rate of the nonparametric estimator defined by (5.6). To motivate the need for an uniform convergence rate, let us consider the problem of estimating an average treatment effects. Suppose in additional to $\{x_i, y_i\}_{i=1}^n$, there is a treatment indicator $t_i$, $t_i = 1$ if individual $i$ has received a treatment, and $t_i = 0$ otherwise. Let $y_i(t_i)$ denote the outcome, and we are interested in finding out the average treatment effect: $\tau = E[y(1) - y(0)]$. Under some conditions including that $t_i$ is independent of the potential outcome conditional on $x_i$, it can be shown that $\tau = E[(t_i - e_i)y_i / var(t_i|x_i)]$, where $e_i = Pr[t_i = 1|x_i]$ is the propensity score. Hahn (1998), and Hirano, Imbens and Ridder (2003) nonparametric estimation of $\tau$ based on nonparametric series method. A typical treatment-type data usually contains many categorical variables, and the conventional nonparametric frequency method can be difficult to apply to this type of data.

One can smooth the categorical variables as described by (5.6) to avoid the sample splitting problem, and therefore, one can estimate $\tau$ by

$$\hat{\tau} = \frac{1}{n} \sum_i \frac{(t_i - \hat{e}_i) y_i}{\hat{e}_i (1 - \hat{e}_i)}, \tag{5.7}$$

where $\hat{e}_i = \sum_j t_i K_{\gamma, z_j, z_i} / \sum_j K_{\gamma, z_j, z_i}$.

To derive the asymptotic distribution of $\hat{\tau}$, one needs the uniform convergence rate of $\hat{e}(x) - e(x)$ to handle the denominator $\hat{e}_i(1 - \hat{e}_i)$. The Theorem V.1 below provides the desired result.

The following conditions will be used in establishing the uniform convergence rate of $\hat{g}(x)$.

(C1) (i) $\{x_i, y_i\}_{i=1}^n$ is a strictly stationary $\alpha$-mixing process with $E[|y_i|^\xi] < \infty$ for some $\xi > 2$. (ii) $g$ and $f$ both satisfy some Lipschitz conditions, i.e., $|G(u) - G(v)| \leq c_2 ||u - v||$ for some finite positive constant $c_2$, where $G = g$ or $f$ ($||.||$ is the

usual Euclidean norm). (iii) Let $\mathcal{S} = \mathcal{S}^c \times \mathcal{S}^d$, where $\mathcal{S}^c$ is a compact subset of $R^q$, $inf_{x\in\mathcal{S}} f(x) \geq \delta > 0$ for some positive $\delta$.

(C2) $w(.)$ is a bounded, symmetric $\nu$th ($\nu \geq 2$) order kernel function, satisfying (i) $\int w(v)dv = 1$, $\int w(v)v^l = 0$ for $l = 1, ..., \nu - 1$, $\int w(v)v^\nu \neq 0$ is finite, (ii) $|w(u) - w(v)| \leq c_1|u - v|$ for some finite positive constant $c_1$, (iii) $\int |w(u)||u|^p du$ is finite for some $p > 2\nu$.

(C3) Define

$T_n = \{n(ln\, n)(ln\, ln\, n)^{1+\delta}\}^{1/\xi}$ for some $0 < \delta < 1$,

$r(n) = [nh_1...h_q/(T_n^2 ln\, n)]^{1/2}$,

$L(n) = \{nT_n^2/[(\sum_{s=1}^q h_s^2)(h_1...h_q) ln\, n]\}^{q/2}$,

$$\psi(n) = \frac{L(n)}{r(n)} \left(\frac{nT_n^2}{ln\, nh_1...h_q}\right)^{1/4} \alpha(r(n)),$$

and

$$\phi_n = \frac{n^{1-2/\xi} h_1...h_q}{ln\, n\{(ln\, n)(ln\, ln\, n)^{1+\delta}\}^{2/\xi}}, \qquad (\xi > 2 \text{ is the same as in (C1)}).$$

Then as $n \to \infty$, $\phi_n \to \infty$, and $\sum_{n=1}^\infty \psi(n) < \infty$.

(C1) contains some standard smoothness and moment conditions. (C2) says that $w(.)$ is a $\nu$th order kernel ($\nu \geq 2$) satisfying a Lipschitz condition. The conditions in (C3) are the same as in Masry (1996) and are needed to handle the dependent nature of the $\alpha$-mixing data, and handle the fact that $y$ ($u$) may not be bounded (using some truncation arguments).

For independent data, $\alpha(r(n)) = 0$ for all $r(n) \geq 1$, (C3) is simplified to: $\phi_n \to \infty$ as $n \to \infty$. Also, if $y$ has finite moments of any order, then $\xi$ can be chosen arbitrarily large, (C3) is implied by $n^{1-\epsilon} h_1...h_q \to \infty$ for some arbitrarily small $\epsilon > 0$.

**Theorem V.1** *Denotes by $H_n = (0, \zeta_n]^{q+r}$, where $\zeta_n$ is a deterministic sequence that*

*goes to* 0 *as* $n \to \infty$ *at a rate slower than any inverse of polynomials in* $n$. *Then Under conditions (C1) to (C3), and assume that* $(h, \lambda) \in H_n$, *we have*

$$sup_{x \in \mathcal{S}} |\hat{g}(x) - g(x)| = O\left( \sum_{s=1}^{q} h_s^{\nu} + \frac{ln\, n}{(nh_1 ... h_q)^{1/2}} \right) \text{ almost surely (a.s.)}$$

The proof of Theorem V.1 is given in the section C.

Hall, Li and Racine (2003) suggest to choose the smoothing parameters $(h, \lambda) = (h_1, ..., h_q, \lambda_1, ..., \lambda_q)$ by the cross-validation method. That is, we choose $(h, \lambda)$ by minimizing the following cross-validation objective function.

$$CV(h, \lambda) = \sum_{i=1}^{n} \left[ y_i - g_{-i}(x_i) \right]^2, \tag{5.8}$$

where $\hat{g}_{-i}(x_i) = \sum_{j \neq i}^{n} y_j K_{\gamma, x_j, x_i} / \sum_{j \neq i}^{n} K_{\gamma, x_j, x_i}$ is the leave-one-out kernel estimator of $g(x_i)$. Let $\hat{h}_s$ and $\hat{\lambda}_s$ denote the cross-validation selected the smoothing parameters. Under the assumption that all the explanatory variable $(x_1^c, ..., x_q^c, x_1^d, ..., x_r^d)$ are relevant regressors, Hall et al further show that,

$$\begin{aligned} \hat{h}_s &= a_s^0 n^{-1/(q+2\nu)} + o_p(n^{-1/(q+2\nu)} \text{ for } s = 1, ..., q; \\ \hat{\lambda}_s &= b_s^0 n^{-2/(q+2\nu)} + o_p(n^{-2/(q+2\nu)} \text{ for } s = 1, ..., r. \end{aligned} \tag{5.9}$$

where $a_s^0$ and $b_s^0$ are some uniquely defined positive constants. If $(h, \lambda)$ are selected by the cross-validation method, then (5.9) implies that $(\hat{h}, \hat{\lambda}) \in H_n = (0, \zeta_n]^{q+r}$ with probability approaching to one as $n \to \infty$. Therefore, one can use $(\hat{h}, \hat{\lambda})$ in computing the semiparametric estimator $\hat{\tau}$ defined in (5.7). In this case one can derive the asymptotic distribution of $\hat{\tau}$ using the following uniform *in probability* convergence rate.

**Theorem V.2** *Let* $\hat{g}(x, \hat{\gamma})$ *denote the nonparametric estimator* $\hat{g}(x)$ *as defined in (5.6) with the smoothing parameter being selected by the cross-validation method* ($\hat{\gamma} =$

$(\hat{h}, \hat{\lambda})$). *Then under the conditions (C1) to (C3) we have*

$$sup_{x\in\mathcal{S}}|\hat{g}(x,\hat{\gamma}) - g(x)| = O_p\left(\frac{ln\, n}{n^{2\nu/(q+2\nu)}}\right).$$

The uniform rate presented in Theorem V.2 is the *in probability* rate rather than the *almost sure* rate. Because we only have the convergence in probability results for $\hat{h}_s$ and $\hat{\lambda}_s$. Nevertheless, this result is sufficient for deriving the asymptotic distributions of many semiparametric estimators (such as $\hat{\tau}$) with cross-validation selected smoothing parameters.

## C. Proofs

We first give a result from Masry (1996) as a lemma below.

**Lemma V.0** *Define* $A_n(x^c) = n^{-1}\sum_{t=1}^n u_t W_{h,x_t^c,x^c}$, *and* $B_n(x^c) = n^{-1}\sum_{t=1}^n [g(x_t^c) - g(x^c)]W_{h,x_t^c,x^c}$. *Let* $D$ *be any compact subset of* $R^q$, *and assume that conditions (C3) holds. Then*

*(i)* $sup_{x^c\in S^c}|\hat{A}_n(x^c)| = O\left\{\left(\frac{ln\, n}{nh_1...h_q}\right)^{1/2}\right\}$ *a.s.,*

*(ii)* $sup_{x^c\in S^c}|\hat{B}_n(x^c)| = O\left\{\left(\frac{ln\, n}{nh_1...h_q}\right)^{1/2}\right\}$ *a.s.*

Proof: (i) is proved in Theorem 2 of Masry (1996, p.579) and (ii) is proved in Theorem 5 of Masry (1996, p.593).

**Proof of Theorem V.1**

We write $\hat{g}(x) - g(x) = [\hat{g}(x) - g(x)]\hat{f}(x)/\hat{f}(x) \equiv \hat{m}(x)/\hat{f}(x)$, where $\hat{m}(x) = [\hat{g}(x) - g(x)]\hat{f}(x)$. We further decompose $\hat{m}(x)$ into $\hat{m}(x) = \hat{m}_1(x) + \hat{m}_2(x)$, where

$$\hat{m}_1(x) = \frac{1}{nh_1...h_q}\sum_{i=1}^n [g(x_i) - g(x)]K_{\gamma,x,x_i}, \text{ and } \hat{m}_2(x) = \frac{1}{nh_1...h_q}\sum_{i=1}^n u_i K_{\gamma,x,x_i}. \tag{5.10}$$

Lemmas A.1 to A.3 below imply that

$$sup_{x\in\mathcal{S}}|\hat{m}(x)| = O\left(\sum_{s=1}^{q} h_s^{\nu} + \sum_{s=1}^{r}\lambda_r + \frac{ln\, n}{(nh_1...h_q)^{1/2}}\right) \text{ (a.s.)} \tag{5.11}$$

Similarly, one can easily show that

$$sup_{x\in\mathcal{S}}|\hat{f}(x) - f(x)| = O\left(\sum_{s=1}^{q} h_s^{\nu} + \sum_{s=1}^{r}\lambda_r + \frac{ln\, n}{(nh_1...h_q)^{1/2}}\right) \text{ (a.s.)} \tag{5.12}$$

(5.11) and (5.12) complete the proof of Theorem V.1.

**Lemma V.1** $E[\hat{m}_1(x)] = O(\sum_{s=1}^{q} h_s^{\nu} + \sum_{s=1}^{r}\lambda_s)$.

Proof: To compute $E[\hat{m}_1(x)]$, we will use the standard change-of-variable argument to lead to a term like $[g(x^c + hv, x^d) - g(x)]f(x^c + hv, x^d) = (gf)(x^c + hv, x^d) - (gf)(x) - g(x)[f(x^c + hv, x^d) - f(x)]$. Then by the Taylor expansion and noting that $W$ is an $\nu$th order kernel function, it gives a term $\sum_{s=1}^{q}[D_s^{\nu}(gf)(x) - g(x)D_s^{\nu}f(x)]$, where $D_s^{\nu}$ is the $\nu$th order partial derivative with respect to $x_s^c$. We also need to define an indicator function that is related to the leading bias term of discrete variables.

$$\mathbf{1}_s(z^d, x^d) = \begin{cases} \mathbf{1}(x_s^d \neq z_s^d)\prod_{t\neq s}\mathbf{1}(x_t^d = z_t^d) & \text{if } x_s^d \text{ is an unordered variable} \\ \mathbf{1}(|x_s^d - z_s^d| = 1)\prod_{t\neq s}\mathbf{1}(x_t^d = z_t^d) & \text{if } x_s^d \text{ is an ordered variable.} \end{cases}$$

Therefore, we have

$$\begin{aligned} E[\hat{m}_1(x)] &= E[(g(x_i) - g(x))K_{\gamma,x_i,x}] = \sum_{z^d\in\mathcal{S}^d}\int [g(z) - g(x)]K_{\gamma,z,x}f(z)dz^c \\ &= \sum_{s=1}^{r}\lambda_s\mathbf{1}_s(z^d, x^d)\int \left[g(z^d, x^c) - g(x)\right]f(z^d, x^c)dx^c \\ &+ \frac{1}{\nu!}\sum_{s=1}^{\nu} h_s^{\nu}\left[(D_s^{\nu}(fg))(x) - g(x)(D_s^{\nu}f)(x)\right]\left[\int w(v)v^{\nu}dv\right] + o_p(\eta_n) \end{aligned} \tag{5.13}$$

where $\eta_n = \sum_{s=1}^{q} h_s^{\nu} + \sum_{s=1}^{r}\lambda_s$, the $o_p(\eta_n)$ term is uniformly in $x \in \mathcal{S}$.

**Lemma V.2** $sup_{x\in\mathcal{S}}|\hat{m}_1(x) - E[\hat{m}_1(x)]| = O(ln\, n/(nh_1...h_q)^{1/2})$.

Proof: Since $\mathcal{S}^d$ has a finite support, and $\mathcal{S}^c$ is a compact set, for any fixed value of $n$, $\mathcal{S}$ can be covered by a finite number $L_n$ (q-dimensional) cubes with centers $x_k = x_{k,n}$ and length $l_n$ $(k = 1, ..., L_n)$. $L_n$ will be chosen sufficiently large so that inside each cube, $x^d$ takes only one fixed value in $\mathcal{S}^d$. Also, it is easy to show that the point-wise variance $var(\hat{m}_1(x)) = O((nh_1...h_q)^{-1})$ so that the smoothing parameters $\lambda_j$'s does not enter the leading term of $var(\hat{m}_1(x))$. Therefore, by exactly the same proof as in the proof of Lemma A.0 (i) one can show that Lemma V.2 holds. A detailed proof of this lemma is lengthy and is omitted here. However, the proof is available from the authors upon request.

**Lemma V.3** $sup_{x\in\mathcal{S}}|\hat{m}_2(x)| = O(ln\, n/(nh_1...h_q)^{1/2})$.

Proof: Lemma V.3 follows from Lemma A.0 (ii) by the same arguments as in Lemma V.2 above.

**Proof of Theorem V.2**

By Hall et al (2004) we know that $\hat{h}_s/h_s^0 - 1 = o_p(1)$ and $\hat{\lambda}_s/\lambda_s^0 - 1 = o_p(1)$. Write $\hat{h}_s = \hat{a}_s n^{-1/(q+2\nu)}$ and $\hat{\lambda}_s = \hat{b}_s n^{-2/(q+2\nu)}$, then we know that $\hat{a}_s \xrightarrow{p} a_s^0$, and $\hat{b}_s \xrightarrow{p} b_s^0$. Denotes by $c = (a_1, ..., a_q, b_1, ..., b_r)$, $\hat{c} = (\hat{a}_1, ..., \hat{a}_q, \hat{b}_1, ..., \hat{b}_r)$, and $c^0 = (a_1^0, ..., a_q^0, b_1^0, ..., b_r^0)$, Then by $||\hat{a}-a^0|| = o_p(1)$ and $||\hat{b}-b^0|| = o_p(1)$, we have $||\hat{c}-c^0|| = o_p(1)$.

Define $\mathcal{C}_\delta = \{c\,|\,||c-c^0|| \le \delta\}$ as a $(q+r)$-dimension ball centered at $c^0$ with radius $\delta$. And define $H_{n,0} = \prod_{s=1}^{q}[a_{1s}n^{-1/(q+2\nu)}, a_{2s}n^{-1/(q+2\nu)}] \prod_{l=1}^{r}[b_{1l}n^{-2/(q+2\nu)}, b_{2l}n^{-2/(q+2\nu)}]$. From $||\hat{c} - c^0|| = o_p(1)$ we know that for all $\delta > 0$,

$$lim_{n\to\infty} Pr[\,\hat{c} \in \mathcal{C}_\delta\,] = 0. \tag{5.14}$$

Since $H_{n,0}$ is compact, by the same arguments as we did in the proof of Theorem V.1, and note that almost sure convergence implies the convergence in probability,

one can easily establish the following uniform rate of convergence in probability.

$$sup_{x\in\mathcal{S},\gamma\in H_{n,0}}|\hat{g}(x,\gamma)-g(x)| = O_p\left(\sum_{s=1}^{q} h_s^{\nu} + \sum_{s=1}^{r}\lambda_s + \frac{ln\, n}{(nh_1...h_q)^{1/2}}\right). \tag{5.15}$$

Now, for all $\delta > 0$ we have

$$sup_{x\in\mathcal{S}}|\hat{g}(x,\hat{\gamma})-g(x)| = sup_{x\in\mathcal{S},\,\hat{c}\in C_\delta}|\hat{g}(x,\hat{\gamma})-g(x)| + sup_{x\in\mathcal{S},\,\hat{c}\notin C_\delta}|\hat{g}(x,\hat{\gamma})-g(x)|. \tag{5.16}$$

We can choose $\delta$ small enough so that $\hat{\gamma}\in H_{n,0}$ will imply that $\hat{c}\in C_\delta$. Therefore, we have

$$\begin{aligned} & sup_{x\in\mathcal{S},\,\hat{c}\in C_\delta}|\hat{g}(x,\hat{\gamma})-g(x)| \\ \leq\ & sup_{x\in\mathcal{S},\gamma\in H_{n,0}}|\hat{g}(x,\gamma)-g(x)| \\ =\ & O_p\left(\sum_{s=1}^{q} h_s^{\nu} + \sum_{s=1}^{r}\lambda_s + \frac{ln\, n}{(nh_1...h_q)^{1/2}}\right) \end{aligned} \tag{5.17}$$

by (5.15).

Let $\eta_{2n} = \sum_{s=1}^{q} h_s^{\nu} + \sum_{s=1}^{r}\lambda_s + \frac{ln\, n}{(nh_1...h_q)^{1/2}}$. Then using (5.14) we have for all $\delta > 0$ that

$$lim_{n\to\infty} Pr[sup_{x\in\mathcal{S},\,\hat{c}\notin C_\delta}\ \eta_{2n}^{-1}|\hat{g}(x,\hat{\gamma})-g(x)| > \delta] \leq lim_{n\to\infty} Pr[\hat{c}\notin C_\delta] = 0. \tag{5.18}$$

Note that (5.18) implies that

$sup_{x\in\mathcal{S},\,\hat{c}\notin C_\delta}|\hat{g}(x,\hat{\gamma})-g(x)| = O_p(\sum_{s=1}^{q} h_s^{\nu} + ln\, n/(nh_1...h_q)^{1/2})$.

Therefore, (5.16) to (5.18) together lead to

$$sup_{x\in\mathcal{S}}|\hat{g}(x,\hat{\gamma})-g(x)| = O_p\left(\sum_{s=1}^{q} h_s^{\nu} + \sum_{s=1}^{r}\lambda_s + \frac{ln\, n}{(nh_1...h_q)^{1/2}}\right)$$

with $h_s = O(n^{-1/(q+2\nu)})$ and $\lambda_s = O(n^{-2/(q+2\nu)})$. This completes the proof of Theorem V.2.

# CHAPTER VI

## SUMMARY

Several topics in the field of nonparametric econometrics are discussed in this dissertation.

In chapter II, we consider the problem of estimating a nonparametric regression model with only categorical regressors. We investigate the theoretical properties of least squares cross-validated smoothing parameter selection, establish the rate of convergence (to zero) of the smoothing parameters for relevant regressors, and show that there is a high probability that the smoothing parameters for irrelevant regressors converge to their upper bound values thereby smoothing out the irrelevant regressors. We also discuss the asymptotic distribution of the resulting estimator. Simulation results and applications of the proposed method in this chapter are also presented.

In chapter III, we consider the problem of estimating a joint distribution defined over a set of discrete variables. We use a smoothing kernel estimator to estimate the joint distribution, allowing for the case in which some of the discrete variables are uniformly distributed, and explicitly address the vector-valued smoothing parameter case due to its practical relevance. We show that the cross-validated smoothing parameters differ in their asymptotic behavior depending on whether a variable is uniformly distributed or not. The asymptotic distribution of the resulting estimator (all-relevant case) is discussed. Simulations results show that our method performs much better than the commonly used frequency methods.

In chapter IV, we consider a k-n-n estimation of regression function with $k$ selected by a cross validation method. We consider both the local constant and local linear cases. In both cases, the convergence rate of of the cross validated $k$ is established. We also discuss the asymptotic distributions of the resulting estimators.

In chapter V, we consider nonparametric estimation of regression functions with mixed categorical and continuous data. The smoothing parameters in the model are selected by a cross-validation method. The uniform convergence rate of the kernel regression function estimator function with weakly dependent data is derived. The uniform convergence rate result is useful in studying the asymptotic distributions of many semi-parametric estimators with mixed type data.

# REFERENCES

Aerts, M., Augustyns, I., and Janssen, P., 1997a. Smoothing sparse multinomial data using local polynomial fitting. Journal of Nonparametric Statistics 8, 127-147.

Aerts, M., Augustyns, I., and Janssen, P., 1997b. Local polynomial estimation of contingency table cell probabilities. Statistics 30, 127-148.

Ahmad, I.A., Cerrito, P.B., 1994. Nonparametric estimation of joint discrete-continuous probability densities with applications. Journal of Statistical Planning and Inference 41, 349-364.

Aitchison, J., Aitken, C.G.G., 1976. Multivariate binary discrimination by the kernel method. Biometrika 63, 413-420.

Bierens, H., 1983. Uniform consistency of kernel estimators of a regression function under generalized conditions. Journal of American Statistical Association 78, 699-707.

Bowman, A.W., 1980. A note on consistency of the kernel method for the analysis of categorical data. Biometrika 67, 682-684.

Bowman, A.W., Hall, P., Titterington, T.D.M., 1984. Cross-validation in nonparametric estimation of probabilities and probability densities. Biometrika 71, 341-351.

Fahrmeir, L., Tutz, G., 1994. Multivariate Statistical Modeling Based on Generalized Linear Models. Springer-Verlag, New York.

Fair, R.C., 1978. A theory of extramarital affairs. Journal of Political Economy 86, 45-61.

Grund, B., 1993. Kernal estimators for cell probabilities. Journal of Multivariate Analysis 46, 283-308.

Grund, B., Hall, P., 1993. On the performance of kernel estimators for high-dimensional sparse binary data. Journal of Multivariate Analysis 44, 321-344.

Hahn, J., 1998. On the role of propensity score in efficient semiparametric estimation of average treatment effects. Econometrica 66, 315-331.

Hall, P., 1981. On nonparametric multivariate binary discrimination. Biometrika 68, 287-294.

Hall, P., Li, Q., Racine, J., 2004. Nonparametric estimation of regression function when there exists irrelevant covariates. Unpublished manuscript, Department of Economics, Texas A&M University, College Station.

Hall, P., Racine, J., Li, Q., 2004. Cross-validation and the estimation of conditional probability densities. Journal of American Statistical Association

99, 1015-1026.

Hall, P., Wand, M., 1988. On nonparametric discrimination using density differences. Biometrika 75, 541-547.

Härdle, W., Hall, P., Marron, J.S., 1988. How far are automatically chosen regression smoothing parameters from their optimum? Journal of the American Statistical Association 83, 86-99.

Härdle, W., Hall, P., Marron, J.S., 1992. Regression smoothing parameters that are not far from their optimum. Journal of the American Statistical Association 87, 227-233.

Härdle, W., Marron, J.S.,1985. Optimal bandwidth selection in nonparametric regression function estimation. The Annals of Statistics 13, 1465-1481.

Hart, J.D., 1997. Nonparametric Smoothing and Lack-of-fit Tests. Springer-Verlag, New York.

Hirano, K., Imbens, G.W., Ridder, G., 2003. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71, 1161-1189.

Izenman, A., 1991. Recent developments in nonparametric density estimation. Journal of the American Statistical Association 413, 205–224.

Kalbfleisch, J.D., Prentice, R.L., 1980. The Statistical Analysis of Failure Time Data. Wiley, New York.

Lee, J., 1990. U-Statistics: Theory and Practice. Marcel Dekker, New York.

Li, Q., Racine, J., 2003. Nonparametric estimation of distributions with categorical and continuous data. Journal of Multivariate Analysis 86, 266-292.

Li, Q., Racine, J., 2004. Cross-validation on local linear estimators. Statistic Sinica 14, 485-512.

Li, Q., Racine, J., 2005. Nonparametric Econometrics: Theory and Practice. Princeton University Press, Princeton, NJ. (manuscript forthcoming)

Liu, Z., Lu, X., 1997. Root-N-Consistent estimation of partially linear model by K-nn method. Econometric Reviews 16, 411-420.

Mack, Y.P., Rosenblatt, M., 1979. Multivariate k-nearest neighbor density estimates. Journal of Multivariate Analysis 9, 1-15.

Masry, E., 1996. Multivariate local polynomial regression for time series: Uniform strong consistency and rates. Journal of Time Series Analysis 17, 571-599.

Racine, J., Li, Q., 2004. Nonparametric estimation of regression functions

with both categorical and continuous data. Journal of Econometrics 119, 99-130.

Scott, D., 1992. Multivariate density estimation: theory, practice, and visualization. John Wiley and Sons, New York.

Simonoff, J.S., 1983. A penalty function approach to smoothing large sparse contingency tables. Annals of Statistics 11, 208-218.

Simonoff, J.S., 1996. Smoothing Methods in Statistics. Springer, New York.

Stute, W., Manteiga, W.G., 1996. NN goodness-of-fit tests for linear models. Journal of Statistical Planning and Inference 53, 75-92.

Titterington, D.M., 1980. A comparative study of kernel-based density estimates for categorical data. Technometrics 22, 259-268.

Tutz, G., 1991. Consistency of cross-validatory choice of smoothing parameters for direct kernel estimates. Computational Statistics Quarterly 4, 295-314.

Wang, M.C., Ryzin, J., 1981. A class of smooth estimators for discrete distributions. Biometrika 68, 301-309.

# VITA

**Desheng Ouyang** Email: deshengouyang@yahoo.com

School of Economics, Shanghai University of Finance and Economics, NO.777, Guoding Rd, Yangpu District, Shanghai, 200433, China

---

**Education**

B.S. (Applied Mathematics), Huazhong University of Science and Technology (August 1985)

M.S. (Applied Mathematics), Huazhong University of Science and Technology (August 1999)

Ph.D. Economics, Texas A&M University (August 2005)

**Fields of Specialization**

Primary: Econometrics and Monetary Theory

Secondary: Game theory